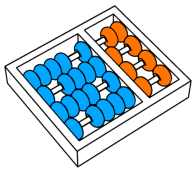
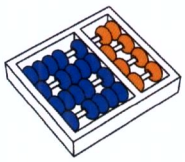


Lin Tzy Li

"A Multimodal Framework for Geocoding Digital Objects"

"*Um Arcabouço Multimodal para Geocodificação de Objetos Digitais*"

CAMPINAS
2014



University of Campinas
Institute of Computing

Universidade Estadual de Campinas
Instituto de Computação

Lin Tzy Li

"A Multimodal Framework for Geocoding Digital Objects"

Supervisor:
Orientador(a): **Prof. Dr. Ricardo da Silva Torres**

"Um Arcabouço Multimodal para Geocodificação de Objetos Digitais"

PhD Thesis presented to the Post Graduate Program of the Institute of Computing of the University of Campinas to obtain a PhD degree in Computer Science.

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Computação da Universidade Estadual de Campinas para obtenção do título de Doutor em Ciência da Computação.

THIS VOLUME CORRESPONDS TO THE FINAL VERSION OF THE THESIS DEFENDED BY LIN TZY LI, UNDER THE SUPERVISION OF PROF. DR. RICARDO DA SILVA TORRES.

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA POR LIN TZY LI, SOB ORIENTAÇÃO DE PROF. DR. RICARDO DA SILVA TORRES.

Ricardo Torres

Supervisor's signature / *Assinatura do Orientador(a)*

CAMPINAS
2014

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Maria Fabiana Bezerra Muller - CRB 8/6162

L612m

Li, Lin Tzy, 1972-
A mutimodal framework for geocoding digital objects / Lin Tzy Li. – Campinas, SP : [s.n.], 2014.

Orientador: Ricardo da Silva Torres.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Fusão de características. 2. Multimodalidade. 3. Sistemas de recuperação da informação. 4. Sistemas de informação geográfica. 5. Bibliotecas digitais. I. Torres, Ricardo da Silva,1977-. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Um arcabouço multimodal para geocodificação de objetos digitais
Palavras-chave em inglês:
Feature fusion
Multimodality
Information storage and retrieval systems
Geographic information systems
Digital libraries
Área de concentração: Ciência da Computação
Titulação: Doutora em Ciência da Computação
Banca examinadora:
Ricardo da Silva Torres [Orientador]
Clodoveu Augusto Davis Junior
Denise Guliato
Luiz Fernando Bittencourt
Ariadne Maria Brito Rizzoni Carvalho
Data de defesa: 27-02-2014
Programa de Pós-Graduação: Ciência da Computação

TERMO DE APROVAÇÃO

Defesa de Tese de Doutorado em Ciência da Computação, apresentada pela Doutoranda Lin Tzy Li, aprovada em 27 de fevereiro de 2014 pela Banca examinadora composta pelos professores doutores:

Prof. Dr. Clodoveu Augusto Davis Junior
DCC / UFMG

Profª. Drª. Denise Guliato
FACOM / UFU

Prof. Dr. Luiz Fernando Bittencourt
IC / UNICAMP

Profª. Drª. Ariadne Maria Brito Rizzoni Carvalho
IC / UNICAMP

Prof. Dr. Ricardo da Silva Torres
IC / UNICAMP

Institute of Computing /*Instituto de Computação*
University of Campinas /*Universidade Estadual de Campinas*

A Multimodal Framework for Geocoding Digital Objects

Lin Tzy Li[1]

February 27, 2014

Examiner Board/*Banca Examinadora*:

- Prof. Dr. Ricardo da Silva Torres (Supervisor/*Orientador*)
- Prof. Dr. Luiz Fernando Bittencourt
  Institute of Computing - UNICAMP
- Prof. Dr. Ariadne Maria Brito Rizzoni Carvalho
  Institute of Computing - UNICAMP
- Prof. Dr. Clodoveu Augusto Davis Junior
  Departamento de Ciência da Computação - UFMG
- Prof. Dr. Denise Guliato
  Faculdade de Computação - UFU
- Prof. Dr. Ilmério Reis da Silva
  Faculdade de Computação - UFU (Substitute/*Suplente*)
- Prof. Dr. Edson Borin
  Institute of Computing - UNICAMP (Substitute/*Suplente*)
- Prof. Dr. Islene Calciolari Garcia
  Institute of Computing - UNICAMP (Substitute/*Suplente*)

[1]Financial support: CAPES scholarship for abroad doctoral internship (process 1385-10-0), Aug./2010–Jul./2011.

Abstract

Geographical information is often enclosed in digital objects (like documents, images, and videos) and its use to support the implementation of different services is of great interest. For example, the implementation of map-based browser services and geographic searches may take advantage of geographic locations associated with digital objects. The implementation of such services, however, demands the use of geocoded data collections.

This work investigates the combination of textual and visual content to geocode digital objects and proposes a rank aggregation framework for multimodal geocoding. Textual and visual information associated with videos and images are used to define ranked lists. These lists are later combined, and the new resulting ranked list is used to define appropriate locations. An architecture that implements the proposed framework is designed in such a way that specific modules for each modality (e.g., textual and visual) can be developed and evolved independently. Another component is a data fusion module responsible for seamlessly combining the ranked lists defined for each modality. Another contribution of this work is related to the proposal of a new effectiveness evaluation measure named Weighted Average Score (WAS). The proposed measure is based on distance scores that are combined to assess how effective a designed/tested approach is, considering its overall geocoding results for a given test dataset.

We validate the proposed framework in two contexts: the MediaEval 2012 Placing Task, whose objective is to automatically assign geographical coordinates to videos; and the task of geocoding photos of buildings from Virginia Tech (VT), USA. In the context of the Placing Task, obtained results show how our multimodal approach improves the geocoding results when compared to methods that rely on a single modality (either textual or visual descriptors). We also show that the proposed multimodal approach yields comparable results to the best submissions to the Placing Task in 2012 using no additional information besides the available development/training data. In the context of the task of geocoding VT building photos, experiments demonstrate that some of the evaluated local descriptors yield effective results. The descriptor selection criteria and their combination improved the results when the knowledge base used has the same characteristics of the test set.

Resumo

Informação geográfica é usualmente encontrada em objetos digitais (como documentos, imagens e vídeos), sendo de grande interesse utilizá-la na implementação de diferentes serviços. Por exemplo, serviços de navegação baseados em mapas e buscas geográficas podem se beneficiar das localizações geográficas associadas a objetos digitais. A implementação destes serviços, no entanto, demanda o uso de coleções de dados geocodificados.

Este trabalho estuda a combinação de conteúdo textual e visual para geocodificar objetos digitais e propõe um arcabouço de agregação de listas para geocodificação multimodal. A informação textual e visual de vídeos e imagens é usada para definir listas ordenadas. Em seguida, elas são combinadas e a nova lista ordenada resultante é usada para definir a localização geográfica de vídeos e imagens. Uma arquitetura que implementa essa proposta foi projetada de modo que módulos específicos para cada modalidade (e.g., textual ou visual) possam ser aperfeiçoados independentemente. Outro componente é o módulo de fusão responsável pela combinação das listas ordenadas definidas por cada modalidade. Outra contribuição deste trabalho é a proposta de uma nova medida de avaliação da efetividade de métodos de geocodificação chamada *Weighted Average Score (WAS)*. Ela é baseada em ponderações de distâncias que permitem avaliar a efetividade de uma abordagem, considerando todos os resultados de geocodificação das amostras de teste.

O arcabouço proposto foi validado em dois contextos: desafio *Placing Task* da iniciativa *MediaEval 2012*, que consiste em atribuir, automaticamente, coordenadas geográficas a vídeos; e geocodificação de fotos de prédios da Virginia Tech (VT), EUA. No contexto do desafio *Placing Task*, os resultados mostram como nossa abordagem melhora a geocodificação em comparação a métodos que apenas contam com uma modalidade (sejam descritores textuais ou visuais). Nós mostramos ainda que a proposta multimodal produziu resultados comparáveis às melhores submissões que também não usavam informações adicionais além daquelas disponibilizadas na base de treinamento. Em relação à geocodificação das fotos de prédios da VT, os experimentos demostraram que alguns dos descritores visuais locais produziram resultados efetivos. A seleção desses descritores e sua combinação melhoraram esses resultados quando a base de conhecimento tinha as mesmas características da base de teste.

Acknowledgements

First, my gratitude to my advisor Dr. Ricardo da S. Torres who has believed, truly supported, and guided me throughout this whole process. Not to mention his belief that I would be able to finish this doctoral work even keeping my full-time job. Then, I thank the professors who accepted to be part of my examiner board either in my research defense and/or my final defense. Their very helpful feedbacks, reviews, and suggestions made this work stronger.

Next, I would like to acknowledge the administrative staff of the institute and university for their backstage work, as well as professors and friends from the Institute of Computing at UNICAMP. Then, I would like to thank colleagues and professor from LIS (Laboratory of Information Systems) and RECOD (Reasoning for Complex Data) research labs and IC-UNICAMP for suggestions, help, friendship, feedbacks, advices, and/or collaborations: Dr. Claudia Bauzer Medeiros, Carla Macário, Rodrigo Senra, Ricardo Panaggio, Nádia Kozievitch, Jefersson Alex dos Santos, Fábio Augusto Faria, Otávio Augusto Bizetto Penatti, Rodrigo Tripodi Calumby, Daniel Carlos Guimarães Pedronette, and Jurandy Almeida. Other people outside UNICAMP who deserve my acknowledgment are Dr. Karla Albuquerque de Vasconcelos Borges (Prodabel) and Patrícia Correia Saraiva (UFAM) for making available some resources they used in their dissertation and thesis for my exploratory work.

I am grateful for CPqD Foundation encouragement and support (as well my friends and colleagues) in some defining moments during my doctoral quest. I especially thank Geovane Cayres Magalhães, Mário Massato Harada, Cláudia Piovesan Macedo, and Márcia Fiorilli Gusson Roscitto.

I also acknowledge the financial support provided by CAPES for my abroad doctoral internship (Sandwich Doctorate scholarship 1385-10-0) at Virginia Tech (VT), USA, and for work presentation in conferences. It enabled further opportunities and collaborations that go beyond the academic sense of those words.

I am thankful to Dr. Edward A. Fox for his advice and time besides accepting me as Ph.D intern student for one year internship in his research lab, the Digital Library Research Library (DLRL) at VT. In DLRL, I was so warmly welcomed that I felt as

part of that family. I was introduced to great people who taught me many multicultural and technical subjects that I can relate to forever. Not to mention many opportunities to grow, learn, collaborate, and make friends. I have enjoyed my time there to learn, hang around, or work with people like Dr. Fox himself, Dr. Eric Hallerman, Dr. Andrea L. Kavanaugh, Dr. Steven D. Sheetz, Dr. Donald Shoemaker, Travis Whalen, Venkat Srinivasan, Seungwon Yang, Sung Hee Park, Sunshin Lee, Spencer J. Lee, Eric Fouh, Susie Marion, Lubna Shihadeh, Jonathan Leidig, Monika Akbar, and Uma Murthy.

As a result of one of many opportunities provided by Dr. Fox at VT, we thank its University Relations (UniRel) for providing access to some of the photographs used in one of our works. Additionally, we would like to thank the Center for Geospatial Information Technology (CGIT) at VT and the GIS Management for VT campus facilities for providing access to the campus building database.

In this doctoral journey, I got to know some Brazilians in Blacksburg as Cristiane and Regis Kopper and Luiza Abruzzi, who have turned into our best friends in US. Regis and Cris have even picked us up at the airport without knowing who my husband and I were when we first arrived there. Not to mention the reunion there with Dr. João Setubal and his wife Silvia, who I have met before in Campinas.

I appreciate very much all the collaborations built in these years and so many shared works in co-authored papers. In most of these submissions, there were online or on-site meetings for last minutes (hours) reviews and wrapping up before the final deadlines. Despite the tiredness, there were some fun and sense of accomplishment in each of these paper/result submissions. Looking forward to more collaborations.

In closing, I would like to express my appreciation to my family and friends support. Specially, from the bottom of my heart, I am deeply thankful to my husband Zanoni Dias, who has always supported me and my doctoral studies besides being so understanding and patient with my weekends and nights sold “a million times” to meet many deadlines. Surely those nights included holidays and the eves of our marriage (and honeymoon), Christmas, and New Year, just to name a few.

"The profoundest distances are never geographical."

- John Fowles

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Since geographic information is involved in people's daily lives, there is a great amount of data about geographical entities on the Web. This information is also often found in digital objects (e.g., documents, images, and videos) of several digital libraries (DLs). The process of associating a geographic location with photos, videos, and documents is called *geocoding*. When a digital object is geocoded, it is related to some place on Earth, and therefore it can be browsed on a map. That opens new opportunities for establishing new relations based on geographic location.

The development of spatially-aware services (e.g., search and browse), on the other hand, demands that digital objects be geocoded or geotagged, i.e., the location of digital objects in terms of their latitude and longitude needs to be defined in advance. *Geocoding* is a common expression used in the Geographic Information Retrieval (GIR) community. Other existing designations like *geotagging* and *georeferencing* usually appear in the multimedia domain [89]. In the Geographic Information System (GIS), georeferencing is a term largely used to refer to a given location where something exists, in a physical space, in terms of a coordinate system (i.e., latitude and longitude).

This work tackles the task of geocoding digital objects. The main motivation for geocoding them is to empower new services with spatial reasoning, such as those that exploit intrinsic relations that exist among geographic entities and are encoded or represented in/by different modalities.

This work proposes a multimodal approach to geocode digital objects. In this work "multi" refers to "more than one" and "modal" to mode, modality or type of data (visual, textual, audio, etc). In Oxford Dictionaries, "mode" is a noun defined as "a way or manner in which something occurs or is experienced, expressed, or done" [92]. Hence, we define "multimodal approach for geocoding digital object" as a method that takes into account multiple types of data (or ways it is expressed) that define or are related to a digital object in order to geocode it.

This chapter presents in Section 1.1 the scenarios that motivate this work. In Section 1.2, we present the research challenges and the objectives of this work. In Section 1.3, in turn, we discuss the hypotheses and the main research questions addressed. Finally, Section 1.4 highlights our main contributions, while Section 1.5 outlines the organization of this work.

1.1 Motivating Scenarios

In this section, we present some scenarios related to the use of geographic information in complex information systems. The first scenario refers to the implementation of map-based browsing services in digital libraries (Section 1.1.1). The second one refers to the use of locations associated with images in a particular digital library (Section 1.1.2).

1.1.1 Map-based Browsing Services

Nowadays, there are many devices with a GPS unit embedded, such as cellphones and cameras, that associate location tags with photos and other published content like Twitter updates, Facebook posts, and other posts in social medias. On the Web, tools like Google Maps[1] and Google Earth[2] are very popular, and partially meet the needs of Web users for geospatial information. By using these tools, users can, for example, find an address on a map, look for directions from one place to another, find nearby points of interest (e.g., restaurants, coffee shops, museums) as pictured in Figure 1.1, and list the nearby streets. Other common queries usually desired by users include finding documents, videos, and photos that refer to a certain location's vicinity. Additionally, large collections of digital objects can be browsed based on the location to which they are related, as shown in Figure 1.2.[3]

A possible usage scenario involves a user looking for information related to a certain place on Earth. To meet this user's need, for example, a digital library (DL), map, or location-based services can provide a browsing service showing the world map on which that user can locate objects of interest found in specific locations. Interaction mechanisms based on map navigation (e.g., zoom and pan) and definition of regions of interest could be employed. At different levels of zoom, diverse information could be shown. For example, at a country level, people could visualize country boundaries, and a list of digital objects found within that region could be shown on a map. As a user zooms in to take a closer look at the map, more detailed information or different summaries could be exhibited

[1] `http://maps.google.com/` *(as of Dec. 2013).*
[2] `http://www.google.com/earth/` *(as of Dec. 2013).*
[3] `http://www.panoramio.com/` *(as of Dec. 2013).*

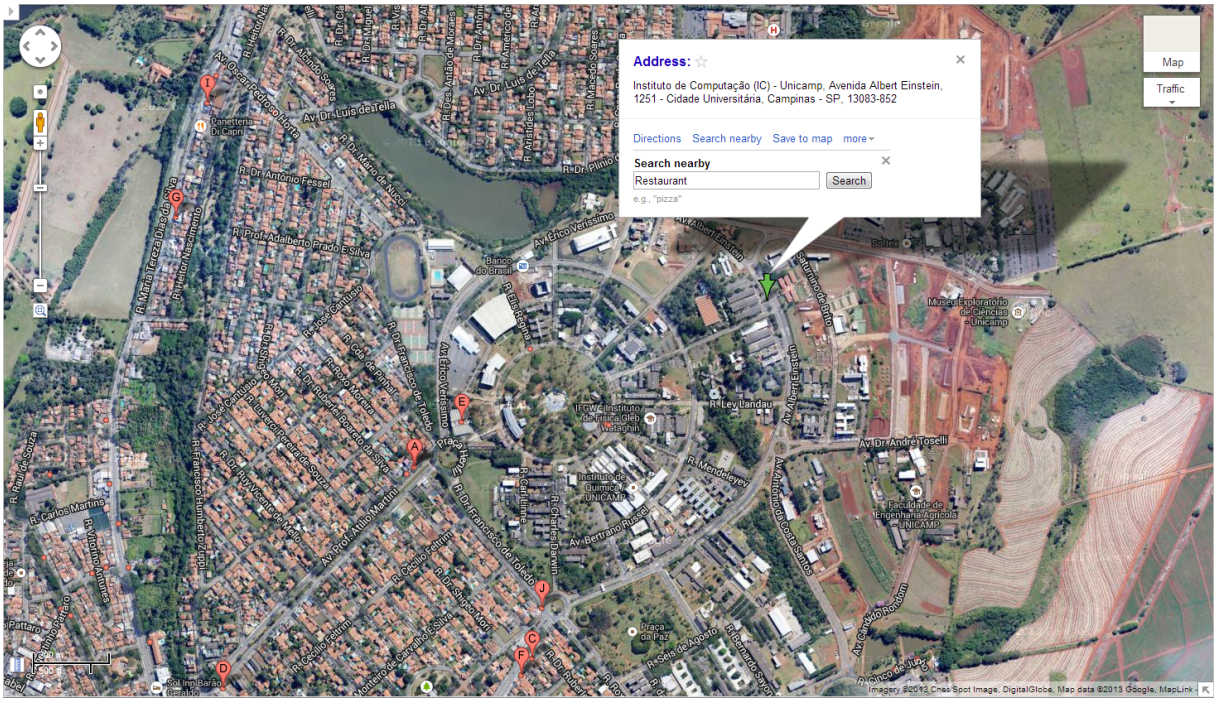


Figure 1.1: Google Maps Search for a point of interest (POI) near by Institute of Computing at University of Campinas (UNICAMP).

as the user is looking at a smaller geographic region. Eventually this user would reach a map zoom level where individual digital objects are shown. By explicitly clicking on then, users could access detailed data related to these resulting objects.

In other scenarios, users might be interested in identifying additional items in a certain place depicted by an image managed by a DL. In this case, a user could use a DL service that can take that image, recognize what and where it refers to, and return all digital objects that are associated with that place. The system interface could show the results pinpointed on a map, helping the user to locate other image digital objects in their spatial context (e.g., cities) and relations (e.g., "are they geographically close?"). For example, let us say that a user uploads a photo from the streets of UNICAMP's campus, then a search service can automatically identify the depicted place, return associated digital objects, and show them on a map along with points of interest.

1.1.2 The CTRnet Digital Library

The CTRnet (Crisis, Tragedy and Recovery Network)[4] [38] DL project collects news and online resources (webpages, public Twitter and Facebook posts) related to natural

[4] `http://www.ctrnet.net/` *(as of Dec. 2013).*

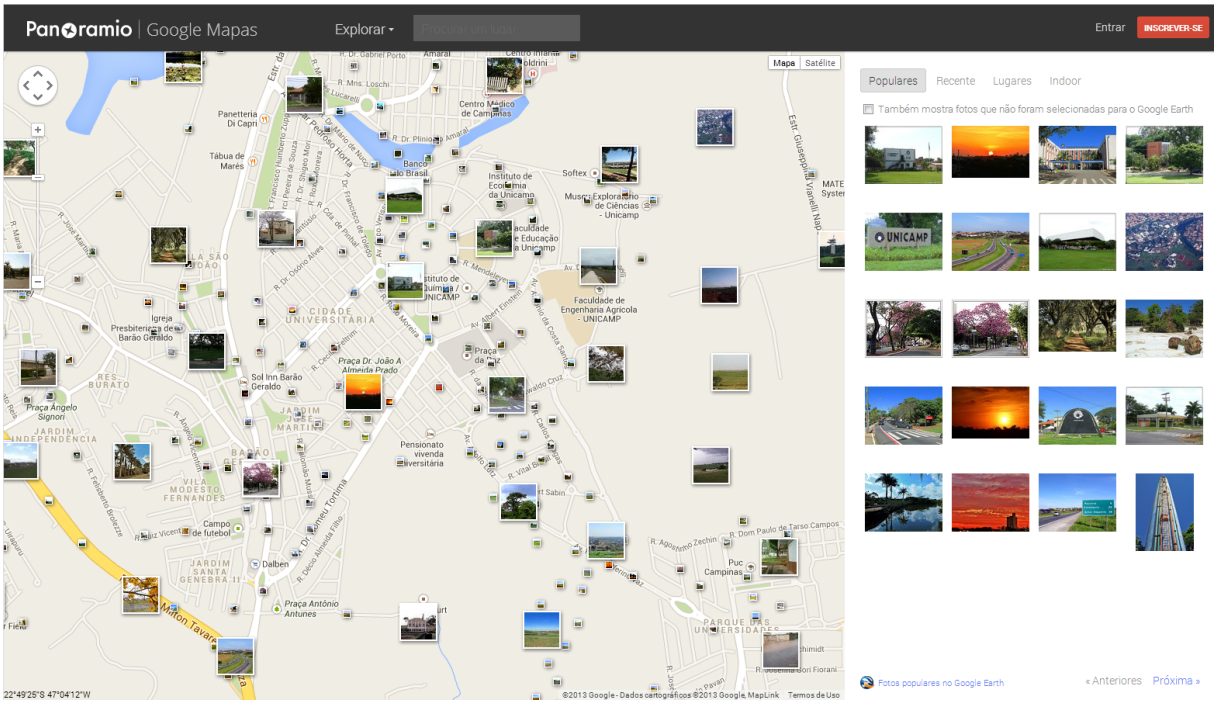


Figure 1.2: Panoramio's browsing service for photos from UNICAMP and its vicinity.

disasters and man-made tragedies [129]. One of the ways to browse through CTRnet collections is through a map interface as shown in Figure 1.3. Purple balloons represent natural disasters, and the blue ones present man-made disasters. Clicking a balloon will open a pop-up window, in which you can visit a corresponding collection in the Internet Archive, or related Wikipedia articles.

In that figure, a user clicked in a place marker near Blacksburg (VA, USA), which opened a pop-up with one of the CTRnet collections associated with that region: the VT (Virginia Tech) April 16 Archive. The cited example is one of the CTRnet collections about the April 16, 2007 school shooting tragedy at Virginia Tech. It refers to the school shooting tragedy *episode* that happened *inside* the VT campus at that very *specific date*. The outcome of that sad event included 32 people murdered before the shooter killed himself.

Now let us consider the following scenario in which geocoded images and collections could be helpful. A journalist has a photo, shown in Figure 1.4, as a query. He might know it is about a school shooting (event), but he would like to know if it appears in the CTRnet collection. If so, he would like to: view the document that used this image, study some facts about that photo if available, and also uncover other related photos so she can reuse them in an essay about that event, about a new similar event, or about the city where that event happened. Therefore meaningful results from the CTRnet DL

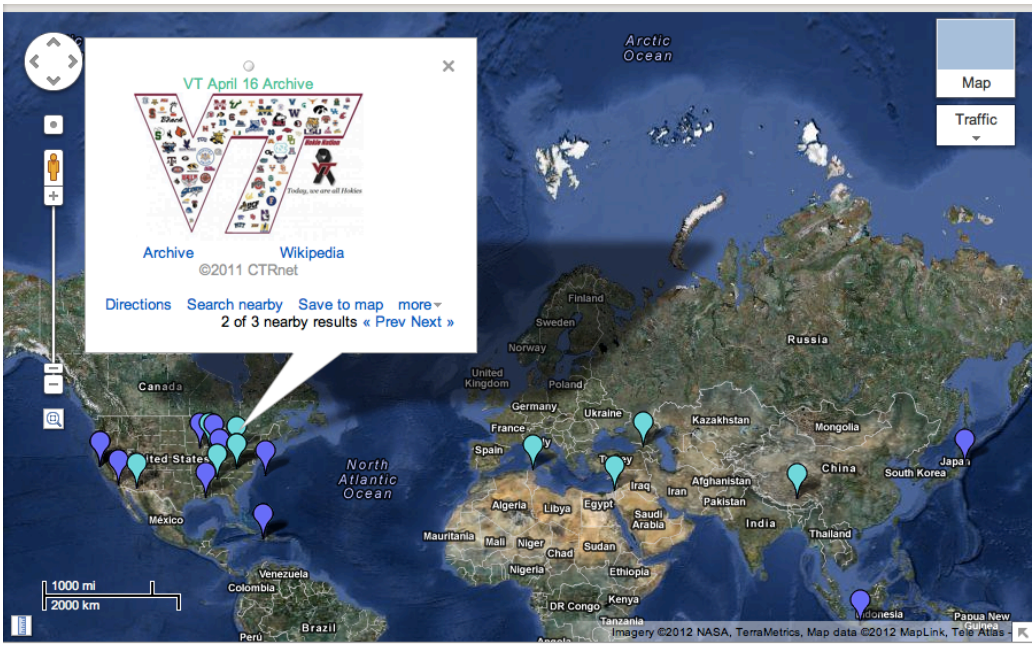


Figure 1.3: CTRnet Collections on Google Maps.

system might be images that look like the query image from these points of view: (a) they are visually similar in terms of scene composition, and (b) there is some relation with Web resources like in the Internet Archive collection, blogs, Twitter, Facebook posts, or news related to the event pictured by that photo. An event is defined by an episode that happened in a certain place and time. Some examples of possible results are shown in Figure 1.5: (1) the archived VT's main webpage from the day after the tragedy; (2) the Wikipedia's webpage about VT's April 16th shooting tragedy; (3) a map of VT main campus highlighting Norris and West Ambler Johnston Hall (where the shootings happened); (4) a photo in The New York Times website reporting the VT's shooting event; (5) a video in VT's Remembrance web site; and (6) a photo of Norris Hall.

One of CTRnet DL aims is to support the creation of map-based browsing and search services based on photo content. The first step towards that is to be able to geocode images. This work describes, in Chapter 4 (Section 4.2), experiments related to the evaluation of image descriptors in image geocoding tasks, and how to combine them to enhance the geocoding results of photos of VT buildings.

1.2 Research Challenges & Objectives

Most of the initiatives for digital object geocoding are based on textual information only [55, 72, 89] (reviewed in Chapter 2). Even some works that claim to be multi-

Figure 1.4: Emergency task force helping the injured in Norris Hall on VT's April 16th shooting tragedy *(by Alan Kim/The Roanoke Times).*

modal in fact reduce the matter to a textual geocoding problem. In the work proposed in [96], for example, other modalities/medias, such as sound/speech, are converted into textual transcripts that are used in text-based geocoding methods.

One problem commonly found on approaches based on textual information relies on the lack of objectivity and completeness, in the sense that the understanding of the visual content of a multimedia object may change according to the experience and perception of each subject. Other challenges include lexical and geographical ambiguities in recognizing place names [75], such as different spelling of the name of a city or country (e.g. Peking and Beijing), indirect references to a place or region (e.g., the Andes), imprecise boundary definitions, and points of interest that remind particular places (e.g., York Properties, Paris Hotel), or places named after a person (e.g., Roosevelt in the state of Utah, USA).

In this scenario, a promising alternative is to use the image/video visual content. The objective is to explore these image/video properties (such as texture, color, and movement) as alternative and complementary cues for geocoding. Furthermore, having multiple (and usually complementary) sources of information for multimedia geocoding also opens the opportunity of using existing fusion approaches to combine them.

There are initiatives in the literature that have proposed some methods to handle the video geocoding problem by exploiting multiple modalities [63, 122] as reviewed in Section 2.2. In these methods, however, the geocoding process consists in the use of *ad hoc* methods (usually one per modality) that are used in a sequential manner to define the location of videos. In these methods, each modality works as a filter that refines the results of previous steps or as a fallback system.

This work aims to investigate the combination of textual and visual content to geocode digital objects, and proposes a flexible multimodal framework for this purpose, so that specific modules for each modality (e.g., textual and visual) can be developed and evolved

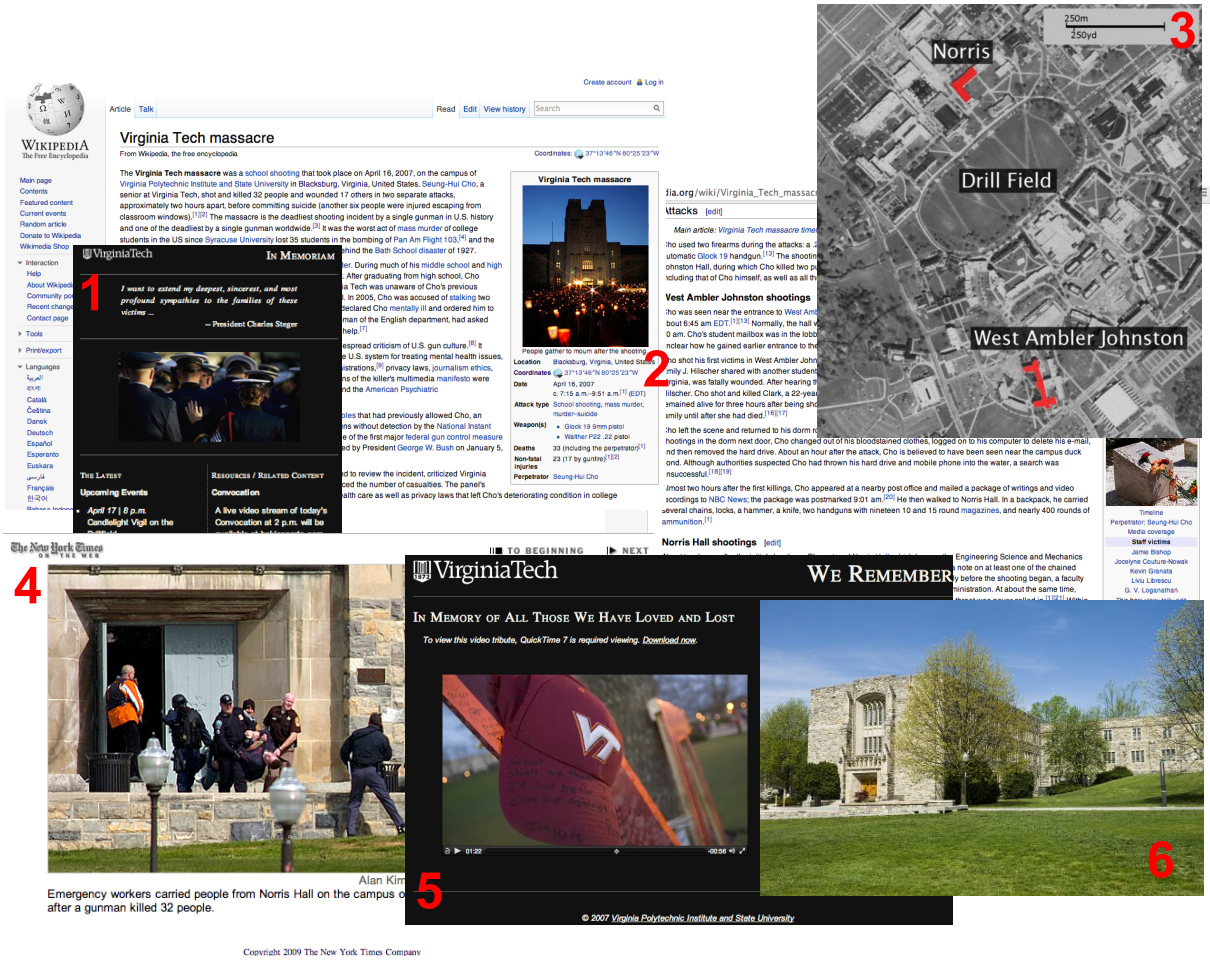


Figure 1.5: Example of results for the query with photo picturing emergency task force at Norris Hall (VT).

independently.

1.3 Hypothesis & Research Questions

This work focuses on geocoding digital objects in order to enabling spatially-aware services, for example geographic searches and map-based browsing facilities. Our proposition is to investigate if combining results based on visual and textual descriptions can improve geocoding results. Thus, our main hypothesis (supported by works in another context) can be enunciated as follows: ***Textual and visual descriptors provide different, but potentially complementary information that can be combined to improve geocoding results.***

One derived hypothesis is that it is possible to integrate independent contributions

developed for each modality (textual and visual) to geocode digital objects. Those contributions are derived from works in the multimedia information retrieval (MIR) and GIR communities. The GIR area deals with the challenges of geoparsing and geocoding textual documents, while MIR handles landscape recognition and scene understanding challenges, using image/video visual content properties.

These hypotheses are associated with the following research questions:

- Does combining visual and textual descriptors enhance geocoding results? A study should be carried out to verify if and when the fusion of various evidences improve the results when compared to geocoding methods that exploit one single modality.
- Which is the role of each feature descriptor (textual and visual) in the geocoding process? Which modality impacts the geocoding results the most? In this context, we deal with textual and visual descriptors, and we need to determine how to choose the more appropriate ones to be used in the fusion approach.
- How do we identify references to places in images or textual documents? Which approaches are appropriate to describe images and textual documents?
- How to combine geocoding results based on visual and textual descriptions? Which fusion strategy is more appropriate for the problem?
- How to define an infrastructure for geocoding based on multimodal descriptions? In this case, we are interested in identifying the components that should compose a multimodal framework that geocodes digital objects. Which contextual information could be used? Which components or tools should be implemented or adapted in order to be used in the geocoding framework?
- How could geocoding strategies be evaluated? Which measures would be appropriate to assess the quality of results provided by geocoding methods?

1.4 Contributions

This work provides contributions in different areas, such as geographic and multimedia information retrieval, as well as digital libraries. In order to enable spatially-aware searching and browsing services, this work is focused on proposing a novel framework for geocoding digital objects that combines different modalities. Although we validate the proposed framework by implementing the geocoding process using textual and visual features of digital objects, we believe that this framework is generic and flexible enough to be extended in order to consider a variety of modalities and implementation strategies, being possibly useful for different applications.

The main contributions of this work are:

- a proposal of a rank aggregation framework for multimodal geocoding of digital objects. It comprises specific modules for each modality (e.g., textual and visual) that can be developed and evolved independently;
- definition of a data fusion module responsible for seamlessly combining ranked lists produced by different modalities;
- partial implementation of the proposed framework using state-of-the-art approaches in the implementation of its components;
- validation of the proposal in the Placing Task at MediaEval benchmark showing the individual performance of textual and visual descriptors, as well as the geocoding results related to the use of fusion approaches;
- validation of the proposed framework in the task of geocoding Virginia Tech (VT) buildings, aiming to enable geo-searching and geo-browsing services related to the VT April 16th collection;
- a new effective evaluation measure to assess the performance of geocoding approaches based on their geocoding results.

1.5 Text Organization

In Chapter 2, we present fundamental concepts related to geographic information; an overview and existing research challenges in the geographic information retrieval area; related work on multimodal retrieval of geographic information; and some initiatives related to the use of geographic information in digital libraries.

Next, in Chapter 3 our proposed framework for geocoding using multimodal descriptions is described and formalized. We also outline its implementation. Finally, we describe a novel effectiveness measure to evaluate geocoding results.

In Chapter 4, we describe experiments carried out in the context of the Placing Task at MediaEval 2012 to validate our proposal for multimodal geocoding. We also describe experiments aiming to geocode photos of VT buildings.

Finally, we conclude this work in Chapter 5, with a summary of our main contributions. We also present possible extensions that could be conducted following some of the main research venues opened by this work.

Chapter 2

Basic Concepts & Related Work

Geographic information is characterized by the existence of an attribute that is related to a localization on Earth, for example a geographic coordinate, or a relationship to some other object whose geographic location is known. It might be a fully complete address (street name, number, and postal code) or even a single reference such as "LaGuardia Airport," which also implicitly relates to the name of the city in which it is located (New York).

An example of a query that most existing *Information Retrieval* systems do not support is: "Which are the webpages of the cities that are neighbors of Blacksburg?" The reason is that spatial operators usually are supported by spatial databases, and those are not integrated with Web search systems. This kind of problem is tackled in the Geographic Information Retrieval (GIR) area, which improves upon information retrieval (IR) by adding the handling of geographic information found in Web documents and queries.

In this chapter, we survey the GIR area. Some of the concepts are related to geospatial (or geographic) information. Others are related to multimodal retrieval, as it integrates with geographic information. A key challenge is recognizing places based on image or video content [49, 62, 72, 89]. Therefore, we also discuss existing initiatives for multimodal geocoding.

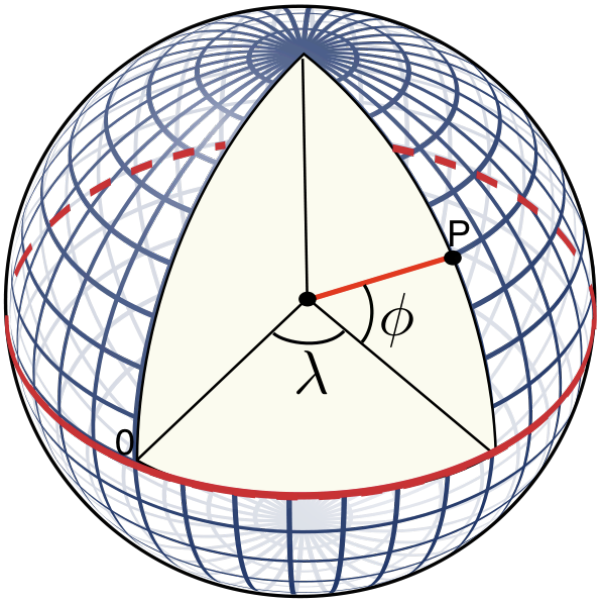
This chapter is organized as follows: in Section 2.1 introductory concepts about geographic information (data format and spatial relations and queries) are explained, followed by a presentation about geographic information retrieval, and discussion on different aspects related to the multimedia retrieval of geographical information and its use in digital libraries; in Section 2.2, related works focused on video geocoding task using multimodal information are summarized; Section 2.3 finalizes this chapter with an overview of the data fusion area and its related works useful for this work.

2.1 Basic Concepts

Fundamental concepts in this field are related to the world of geographic information, which is at the heart of a Geographic Information System (GIS).

A geographic entity/object (e.g., city, country, lake, etc.) can be located on Earth because of the use of a coordinate system. Given an (x, y) coordinate point, x representing a horizontal position and y a vertical one, we can distinguish from other points in the coordinate system space.

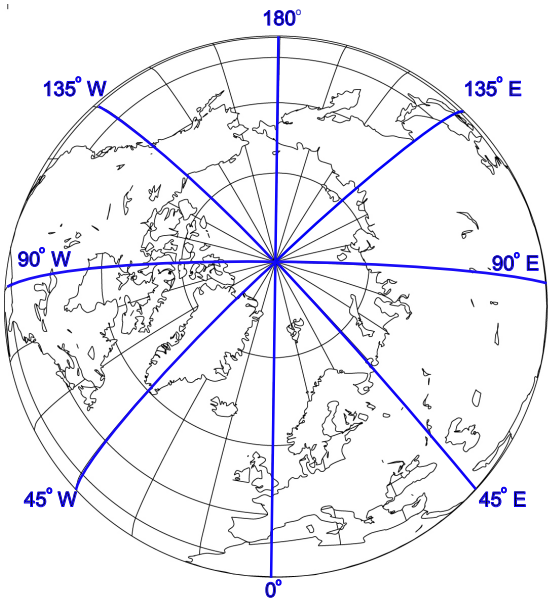
The most popular and ancient coordinate system to locate points on Earth is the geographic coordinate system; every point is at the intersection of an meridian (longitude) and a parallel (latitude). The coordinates are measured in degrees in relation to the center of the globe that represents the Earth (Figure 2.1).



by Peter Mercator. *https://commons.wikimedia.org/wiki/File:Latitude_and_longitude_graticule_on_a_sphere.svg* (as of Nov. 2013).

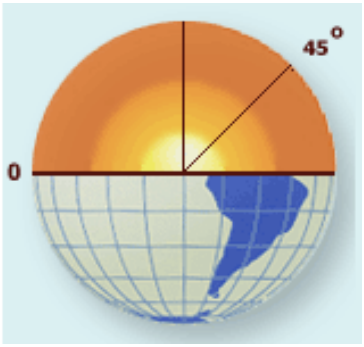
Figure 2.1: Cutaway view of Earth: P is located at latitude $\phi°N$ and longitude $\lambda°E$.

A meridian is an imaginary arc on the Earth's nearly spherical surface that is drawn from the North Pole to the South Pole. The meridians are vertical lines of longitude. Longitude 0 degrees is called the Prime Meridian, or the Greenwich Meridian, that passes through the Greenwich Observatory in England. To the east of the Prime Meridian there are 180 degrees of longitude and to the west another 180 degrees. The east and west directions can be replaced by positive and negative signs, respectively. For example, $105°W$ is equal to -105°. Figure 2.2a shows a schematic view from the North Pole to illustrate longitude lines and how they are drawn.



Built on clip art provided by *www.freeusandworldmaps.com*

(a) North pole globe: Longitude lines (radiuses) and latitude lines (concentric circles). Blue/Darker lines spot some longitude lines.



Source: *nationalatlas.gov*

(b) Cutaway view of Earth showing latitude $45°N$.

Figure 2.2: Longitude and latitude concepts.

On the other hand, the Equator is an imaginary line around the Earth that divides it into two hemispheres (North and South). It marks the 0 degree latitude line. All other latitude lines are parallel and equidistant from each other; thus the latitude lines are known as parallels. There are 90 degrees of latitude to the north and to the south. Parallels above (north of) the Equator are represented as positive degrees and conversely those below (south) appear as negative degrees. For example, $45°N$ is equal to $+45°$. Figure 2.2b shows that latitude is the angle measured from the center of the sphere.

2.1.1 Raster & Vector Data

There are essentially two kinds of geographic data used in GIS:

Raster data comes from satellite images or digital aerial photos, for example, and it is stored as a matrix of cells (or pixels) arranged in rows and columns. Each cell stores some data value, which is the target information. Raster data will have an origin point that will serve as reference for other cells' relative position. Based on its raster coordinate system, a GIS is able to calculate the real-world location for every cell in a raster. This kind of data is useful for continuous data where contours or well-defined shapes are not necessary. Some common image formats in raster are,

for example, JPG, PNG, BMP, GeoTIFF (which embeds geo location in a special tag), etc.

Vector data represents geographic objects like rivers, city boundaries, and houses as basic geometric forms of lines, polygons, and points. As we have seen previously, geographic objects have coordinates (such as latitude and longitude) that associate them with a location on Earth. A point is defined by a coordinate, a line by two coordinates, and a polygon by three or more. Examples of popular vector format files are SVG, DXF, and shapefiles (SHP).

Examples of these two data formats are shown in Figure 2.3.

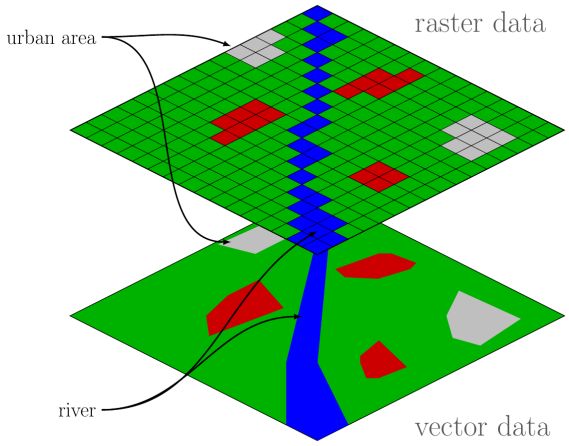


Figure 2.3: Difference between vector and raster data.

Some current database management systems (DBMSs) support storing geographic vector data and provide special operators and functions to query them, e.g., MySQL and PostgreSQL (with PostGIS extension).

2.1.2 Spatial Relationships and Queries

Spatial relationships refer to relative positions between objects in space and they can be classified as [9]:

Topological: this kind of relationship indicates connections between objects such as adjacent to, containing, or is contained, but it does not include measurement or direction. Egenhofer [32] classifies the topological relationships between two dimensional objects as: disjoint, meet, overlap, covers, contains, equal, covered by, and

inside. Clementini et al. [23] summarize them as disjoint, inside, touch, cross, and overlap (Figure 2.4);

Metric: this relationship expresses quantitative measurable attributes like area, distance, length, and perimeter;

Directional: this relationship is used to express orientation such as cardinal points (e.g., North, South, East, and West), as well as order or position like ahead, above, and under.

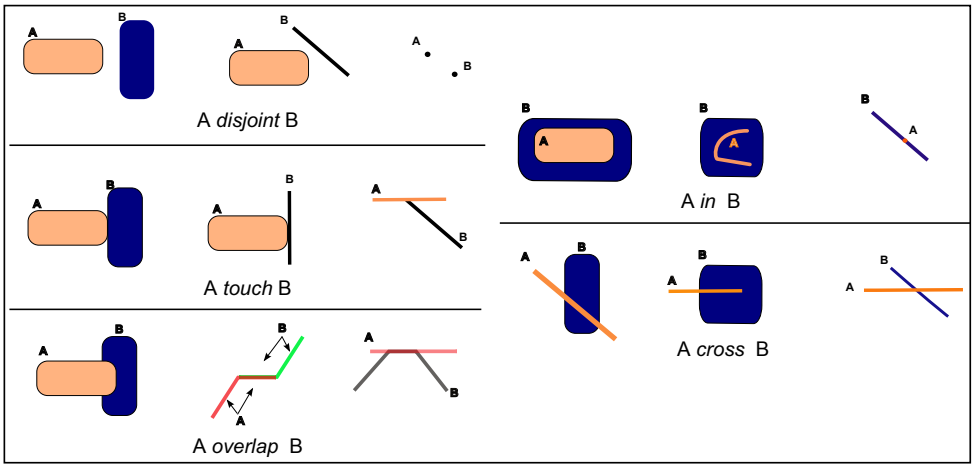


Figure 2.4: Examples of topographic relationship: disjoint, touch, overlap, in (inside), and cross (based on Figure 5.3 in [17]).

These concepts lead us to spatial queries, also known as geographic queries, which express spatial relationships between two objects in a very well defined space, either with or without geographic coordinates such as latitude and longitude. Examples of spatial queries can be organized as [74]:

About a given point inside a coordinate system, like "What can be found at the point given by the following latitude and longitude: 37.228, -80.423?"

About a region, where you are interested in something inside it, e.g., "In which state or region is the Grand Canyon located?"

About distance and a buffer zone: This is illustrated by queries like "Which are the cities 50 miles from the boundaries of Blacksburg?"

Path involves searching along a structured network comprised of connected lines, such as electric lines and networks, water or gas pipes, or transportation lines. Examples

include the shortest path between two points in a network and even a more complicated query like "What is the fastest path from Blacksburg to Washington, D.C.?", which involves distinct variables such as distance, direction, and even time;

Multimedia, when a query requires a variety of information types (e.g., text, image, and geographic), e.g., "In which rivers can we find fishes similar to a given picture, and that are from the darter family?"

2.1.3 Geographic Information Retrieval

Geographical Information Retrieval (GIR) is an area concerned with challenges such as recognizing, querying, retrieving, and indexing geographical information. It combines research in databases, human-computer interaction (HCI), geographic information systems (GIS), indexing, information retrieval (IR), and georeferenced information browsing [74], as well as visualization of information on maps. According to Jones & Purves [55], GIR aims to improve information retrieval centered on geographic information in non-structured documents such as those found in the Web.

Two important concepts of this area are geoparsing and geocoding. Geoparsing is a process of recognizing references with locations inside documents, while ignoring false references (e.g., a place name that is also the name of an organization or person), while geocoding is a process of associating a document with some specific latitude and longitude based on locations recognized by geoparsing [53, 55]. Thus, geocoding consists in mapping a document to a location on Earth. For example, based on where its content refers to, we can assign a latitude and longitude to a document, so later a user can retrieve it based on geographical queries (e.g.,"Give me all documents that refer to parks in the Blacksburg vicinity.").

In the following, we discuss the importance of handling geographic information on the Web, and present a typical GIR architecture. This architecture will serve as a baseline to discuss the main concepts related to GIR – geoparsing and geocoding – as well as existing research challenges.

Geographic Information on the Web

As was introduced earlier, traditional search services are based on keyword matching and do not consider that keywords might represent geographical entities which are spatially related to each other. Yet, even though these relationships have not been explicitly used in a query, they are potentially relevant to users [54].

For example, typing "cities which are neighbors of Campinas" in Brazil into Google search will return webpages with the typed in terms (Figure 2.5). However, that query en-

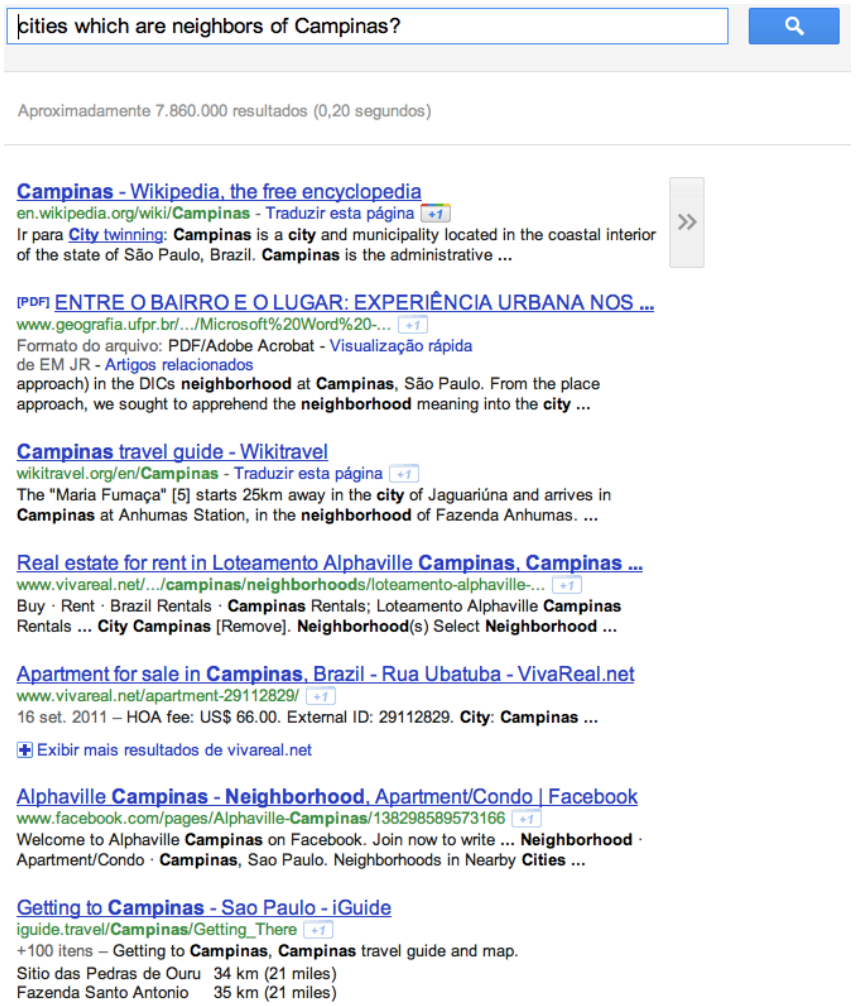


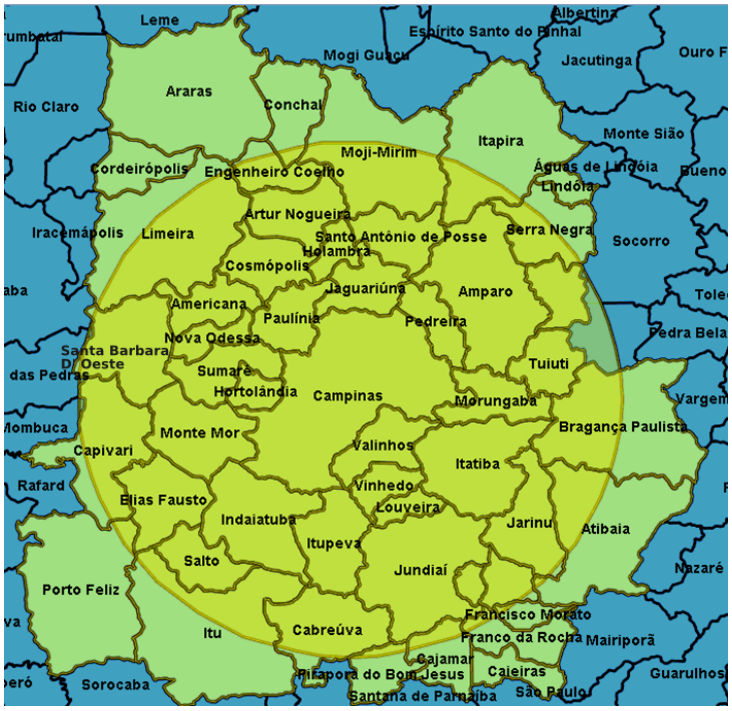
Figure 2.5: Google Search result for neighbors of Campinas, Brazil.

compasses a geographic query (neighbors) meaning that all (ten) cities that share boundaries with Campinas (Figure 2.6), although not mentioned, should be returned in the result set as well.

The difficulty in processing this kind of query comes from the need to combine traditional queries executed on Web search mechanisms with spatial operators usually implemented in spatial databases.

From 2002 to 2005, a project led by Cardiff University (UK) called SPIRIT – *Spatially-Aware Information Retrieval on the Internet* – aimed to develop a spatially-aware Web search engine. They tackled problems such as assigning a footprint to webpages, indexing, searching, ranking, and also user interfaces for geographic queries and results [54, 111].

In [81], users were invited to use Web tools to perform tasks related to the search of geographic entities. All proposed tasks included at least one kind of spatial relation-



Vectorial map provided by IBGE and loaded in QGIS.

Figure 2.6: Campinas neighborhood and cities within 50 km.

ship. Obtained results indicate that: (a) there is a tendency of users to break down the geographic searches into two or more steps when using the current multi-purpose search tools; (b) the geographic relationship aspect of a spatial query on the Web is solved by users either visually inspecting the location on a Web map or by taking advantage of Web tools such as Wikipedia; (c) geographic queries involving well-known or popular objects such as hotel locations in a city are solved easily by a single text-based search.

Let us take as an example one query used in that experiment: "Search for webpages of cities in the neighborhood of Curitiba, Brazil." To perform that task, users switched between keyword-based search, Web mapping, and encyclopedia tools.

It is not enough to send the city's name (e.g., Curitiba) and a term referring to a geographical relationship to a search tool, because it will just match the keywords with the textual content of documents. For those who are used to geographic queries, it is common to rewrite the query into a form which the search tools can use to retrieve relevant results.

To sum up, as a rule, the task of answering a complex geographical query is broken down into two steps. Keeping in mind the example we mentioned earlier, in the first step, the user processes the geographical part of the query by using a keyword search Web tool or Web mapping tool. This step consists of, for example:

- using prior knowledge to associate a city with a region;
- visiting previously known websites (e.g., Wikipedia). From these webpages, users could find nearby cities, or the distance between cities. In this case, users go first to Wikipedia to find the city of interest and then create a list of candidate cities;
- submitting other words to the search tool in order to return the list of cities. This is the case in which the user first searches using the phrase "Curitiba metropolitan cities" to get the list of cities in the Curitiba metropolitan area, thus resolving the part of the query that refers to neighborhood of Curitiba;
- using a map service to locate/find a city used as reference, visually inspecting a map, and manually creating a list of cities that satisfy the target geographical relationship. An example involves going first to a mapping tool like Google Earth or Google Maps, finding the city of interest (e.g., Curitiba), and then by visually picking the neighboring cities.

Finally, the second step involves searching for each city listed in the first step by:

- submitting the city name as keyword to the search tool to find the webpage of that specific city;
- reaching the city page by using a previously known URL naming pattern (e.g., the URL of the home page of a city in Brazil is formed by *www.*<*city name*>*.*<*acronym*>*. gov.br*, where <*city name*> is the city name and <acronym> is the acronym of its state, so for Curitiba, a city in the state of Paraná (PR), its URL is `http://www.curitiba.pr.gov.br`.

There were some cases where a user just relied on map tools, like Google Earth, that show the cities and their amenities and points of interest (POIs) (e.g., hotels and subway stations) on a map to answer the question posed by the task (e.g., "Barcelona's hotels that are near subway stations"). In fact, using these tools, the geographic relationship is resolved by the user, who infers and inspects it visually on the map. Therefore there is no automatic list; in this case the user is in charge of building the list manually.

Incorporating geographic relationships in Web searches is not supported yet; as seen in this study, they are mostly processed by the user first. This could be explained by their inherent complexity, which is worsened by their imprecision and subjectivity. Thus, some geo-related concepts like near or south might depend on the user's search context [56, 134].

For some specific objects (e.g., hotels and city names) and relationships, when geographic terms (e.g., near) can be found on some webpages, the use of current search tools is quite straightforward, as is illustrated by queries like: "webpages of Barcelona's hotels

which are near to subway stations." Such success is explained by those objects' search popularity [51, 56, 114]; webpages that contain those keywords thus can be retrieved by popular keyword-based search tools.

The ideal Web search tool for geographic queries should be able to process the geographical relationships and retrieve all the relevant results on the Web that match the users' intention expressed by their query. This kind of query is common in a GIS (geographical information system) which works with structured data. Hence, there is need for investigation of strategies to integrate these technologies, so that Web queries that include this kind of feature can be easily processed, without undue user frustration or effort. One perspective is that geocoding webpages can help, but the challenge is how to do that in light of the volume of data on the Web and the high level of ambiguity common in such queries.

GIR Architecture

As shown in Figure 2.7, a GIR system can be divided into three layers: presentation, processing, and data.

The main modules for each layer are presented next.

Presentation Layer

Query Input, Result Presentation, and Feedback Presentation. These modules deal with HCI aspects: query input by the user, presentation of results returned by the system, and user feedback about the results. They forward data to lower-level modules aiming at improving obtained results.

Processing Layer

Geoparsing is a module responsible for recognizing references to geographic entities in a digital object and for disambiguating them based on their content, geo-ontologies [115], and semantic databases;

Geocoding is a module that takes care of associating appropriate geographic coordinates with a digital object, which can be one or more geographic points or even a geographic region;

Query Processing is responsible for interpreting and processing the input query; besides, it handles subsequent interactions aiming to refine results;

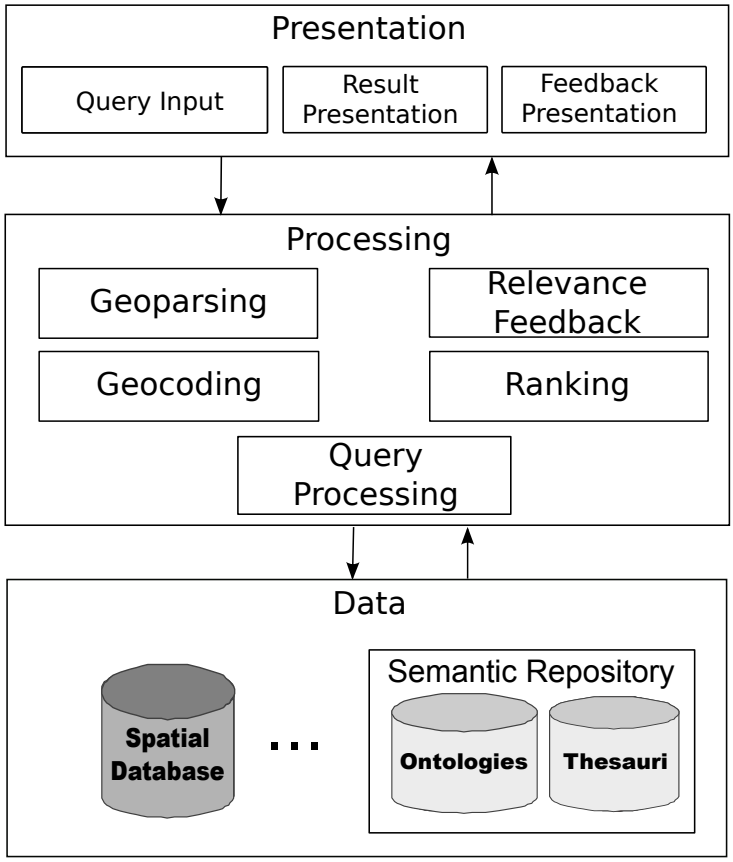


Figure 2.7: Architecture of a GIR system.

Relevance Feedback aims at making improvements in retrieval and/or ranking algorithms by taking into account user assessment of results previously returned. The objective is to return better results through one or a series of interactions;

Ranking is a module in charge of ordering the results according to an estimation of their relevance to the user query.

Data Layer

Semantic Repositories store place names and define the related concepts and how they are organized and related to each other. The objective is to support geoparsing and geocoding tasks. The geographic knowledge can be represented by ontologies; these are called geo-ontologies [54, 115] or geographic ontologies [9]. More information about places can be found in gazetteers and/or thesauri, and even from previously geocoded webpages. A gazetteer is a dictionary of geographic names consisting of a

name and its variants, that place's location, and its category (populated place, school, farm, hotel, lake, etc.). An example of a gazetteer is Geonames.[1] A thesaurus is a list of structured and defined terms formally organized and with concept relations clearly drawn [13], which is what distinguishes thesauri from gazetteers. For example, the *Getty Thesaurus of Geographic Names*[2] organizes places/location based on their spatial relation and administrative area, gives their geographic coordinates and all other names a place has, and supports places with similar names assisted by ontologies [126];

Spatial Database: in addition to what a regular database management system (DBMS) offers, a spatial database also stores and provides spatial operations and queries over stored geographic objects. These objects can be stored using points, lines, or polygons in a given coordinate system. Spatial indexes are built to later speed up spatial/geographic queries (Section 2.1.2). Examples of geographic objects that can be stored are: shapes representing boundaries of a state, city, or country; other shapes representing a specific area on Earth.

Geoparsing & Geocoding

Items in a collection can be associated with one or more regions on Earth, i.e., we can determine their *footprint* [41]. Jones [53, 54] defines **geocoding** as the act of associating a *footprint* with a geographic reference. Recognizing geographic references inside a document is called **geoparsing**, as we introduced previously. Additionally, Jones [53] defines geotagging as a composition of geoparsing and geocoding process.

In GIR, a collection of documents that refer directly or indirectly to a place needs to have their footprint identified and thus be indexed spatially. That is, documents should be geoparsed and then geocoded.

Geoparsing should be able to identify and disambiguate a place name appearing in a document and rule out false references to it. It can be seen as a particular case of name entity recognition (NER), which identifies expressions in text and classifies them as person, event, organization, etc. [5]. In fact, NER is one of the techniques used in geoparsing.

There are some challenges involved in recognizing place references and associating them with their coordinates [75, 126], such as:

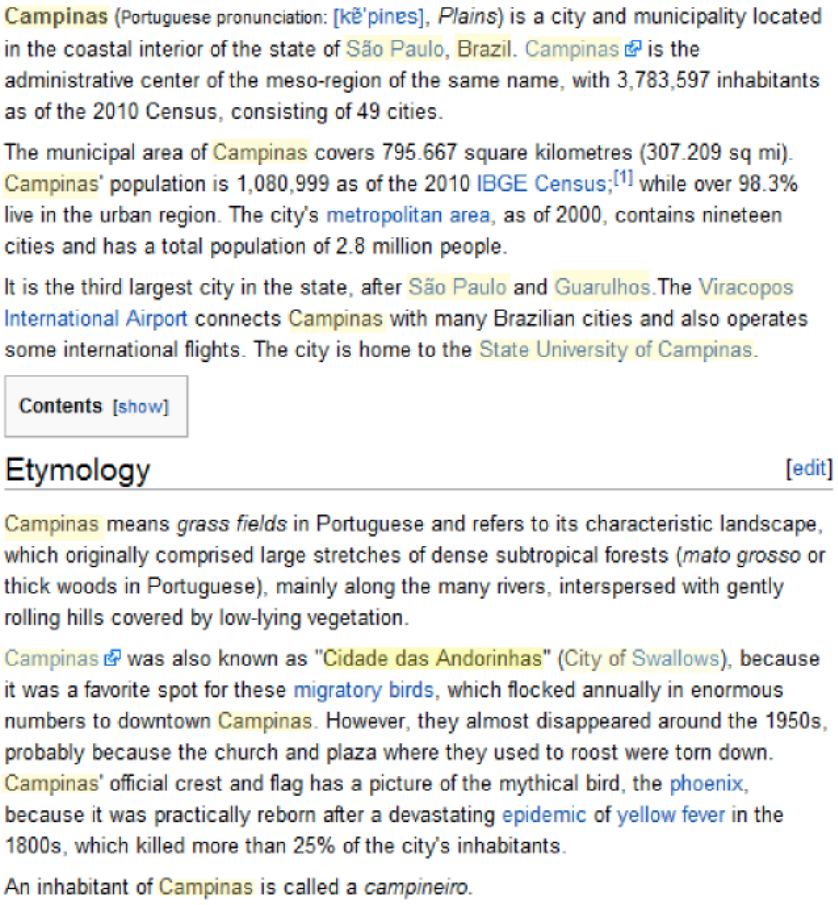
- Homonyms for places and persons: For example, London is city name in UK,

[1] `http://www.geonames.org/` *(as of Dec. 2013).*
[2] `http://www.getty.edu/research/tools/vocabularies/tgn/` *(as of Dec. 2013).*

Canada, and the USA. Additionally, Luis Eduardo Magalhães is a Brazilian politician who also has an airport, square, and city named after him. These examples are called geo/geo and geo/non-geo ambiguities, respectively [4];

- Descriptive place names change according to the historical context, culture, and customs that are in place, when a textual form is produced. For example, locating "North of the Russian capital" on a map would be difficult because the location of the capital of Russia has changed several times;
- Names of places change over time. For example, St. Petersburg, once Russia's capital, was called Petrograd (1914-1924) and Leningrad (1924 - 1991);
- Geographic boundaries change over time. For example, Germany had different boundaries over its history;
- Boundaries cannot always be clearly defined, for example in a conflict zone (e.g., Syrian-Turkish territory dispute);
- Names assigned to regions can refer to an area, rather than a well defined place, e.g., Southern California, or the Andes (in South America);
- Different names may refer to the same geographic entity, whether by error, language variations, or the legal existence of more than one valid way of writing it. For example, both Peking and Beijing refer to the capital of China, and Deutschland is commonly used to refer to Germany;
- Ambiguities arise due to different ways of describing a place, e.g., pseudonyms or expressions used in a specific context. For example, Saint Petersburg is called Piter by locals, while New York City is also referred to as the Big Apple. In Brazil the city of São José do Rio Preto, in São Paulo state, sometimes is called Rio Preto by locals, but, in another Brazilian state (Minas Gerais), Rio Preto is the official name of a different city;
- Names of famous buildings can lend their names to states; thus, New York is sometimes called the Empire State, after that tall building in NYC;
- Indirect references, such as to a road, like the Blue Ridge Parkway, may bring to mind both a region and event, e.g., due to a scenic drive in southwest Virginia and northwest North Carolina;
- Imprecise references such as "100 km from Blacksburg" can refer to some point within 95 and 110 km. In another example, "South of Campinas" might include not just south, but also southeast and southwest locations.

In summary, there are problems related to ambiguity (geo/geo and geo/non-geo), variations (synonyms), indirect references, imprecision, and fuzziness, etc.



Campinas (Portuguese pronunciation: [kẽ'pinɐs], *Plains*) is a city and municipality located in the coastal interior of the state of São Paulo, Brazil. Campinas is the administrative center of the meso-region of the same name, with 3,783,597 inhabitants as of the 2010 Census, consisting of 49 cities.

The municipal area of Campinas covers 795.667 square kilometres (307.209 sq mi). Campinas' population is 1,080,999 as of the 2010 IBGE Census;[1] while over 98.3% live in the urban region. The city's metropolitan area, as of 2000, contains nineteen cities and has a total population of 2.8 million people.

It is the third largest city in the state, after São Paulo and Guarulhos.The Viracopos International Airport connects Campinas with many Brazilian cities and also operates some international flights. The city is home to the State University of Campinas.

Contents [show]

Etymology [edit]

Campinas means *grass fields* in Portuguese and refers to its characteristic landscape, which originally comprised large stretches of dense subtropical forests (*mato grosso* or thick woods in Portuguese), mainly along the many rivers, interspersed with gently rolling hills covered by low-lying vegetation.

Campinas was also known as "Cidade das Andorinhas" (City of Swallows), because it was a favorite spot for these migratory birds, which flocked annually in enormous numbers to downtown Campinas. However, they almost disappeared around the 1950s, probably because the church and plaza where they used to roost were torn down. Campinas' official crest and flag has a picture of the mythical bird, the phoenix, because it was practically reborn after a devastating epidemic of yellow fever in the 1800s, which killed more than 25% of the city's inhabitants.

An inhabitant of Campinas is called a *campineiro*.

Figure 2.8: Geoparsing example: place names recognized in this extract of Wikipedia's page about Campinas (as of Nov. 3rd, 2011).

Figure 2.8 illustrates an accurate geoparsing. The names highlighted should be identified when geoparsing is applied to the shown text. Figure 2.9, on the other hand, illustrates geoparsing with errors, i.e., both true and false references [53]. In this case, false geographic references include personal names (Smedes York, Jack London), business names (Darchester Hotel, York Properties), and common words that are also places (bath, battle, derby, over, well). A possible strategy to distinguish between false and true references is to look for patterns and context [53]:

- For personal names, such as Jack London and Mr. York, the pattern is a first name or title followed by a location name;

- Business names like Paris Hotel have a location word preceded or followed by a business type;
- Detecting a spatial preposition helps to validate a possible location; examples include in, near, south of, outside, etc., as in "I lived in Blacksburg";
- A street name can be distinguished from a city (e.g., Oxford Street) by verifying if its pattern is a location name followed by a road type, for example.

JACK HAGEL, Staff Writer
Redevelopment of the World Trade Center site in New York is getting some input from a Raleigh real-estate maven.

York Properties President Smedes York was chairman of an Urban Land Institute panel at the World Trade Center and Lower Manhattan Summit last month.

The group heard presentations on how the area surrounding the site of the Sept. 11, 2001, terrorist attacks should be redeveloped. It suggested retail be a central focus for developers. The institute will issue a report based on the recommendations before the end of the year.

York was chairman of the Urban Land Institute, a Washington nonprofit organization, from 1989 to 1991. His dad, J.W. "Willie" York, joined the Urban Land Institute in 1947. That's where he met J.C. Nichols, the developer of Country Club Plaza in Kansas City, Mo. -- the center that inspired Willie York to build Raleigh's Cameron Village, the Southeast's first shopping center.

Figure 2.9: True and false references in geoparsing [53].

Some of the tools used in geoparsing and geocoding, to help to detect and disambiguate place references, are geo-ontologies [54], gazetteers, and thesauri. Even Wikipedia has been used to enrich the knowledge base for geoparsing [29]. Thus, places can be referenced using an urban address, postal code, or the area code of a phone number [11].

According to the earlier discussion of modules, geocoding is a process of associating a document/digital object with some specific latitude and longitude, based on location references recognized by geoparsing. In fact, a document can be associated with one or more geographic objects, which in turn can be represented using a point, line, or polygon. Therefore, it is better to define geocoding as the process of associating a digital object with one or more geographic objects instead of just a point on Earth.

As we can observe in Figure 2.8, a set of place names can be recognized in a document. Therefore, one geocoding challenge is to determine which coordinates should be associated

with a given document. This often requires the disambiguation of locations [5]. As was illustrated above, often the same name is used for different geographic locations (referent ambiguity), or the same location is described by different names (reference ambiguity).

The geographic knowledge required for this task is provided by a geo-ontology, supporting structuring, representation, and storage. It includes all suitable data types: place name, place type (city, state, country, etc.), footprint, relation (e.g., containment, adjacency) to other place names, population, historic names and dates, activities, etc. Given a set of geoparsed names, the geocoding process finds the corresponding matches in the geo-ontology. Then, based on related information, a decision can be made regarding a location to contribute, along with a document footprint: keeping, merging, creating, or discarding. Related information could specify how two extracted locations in a document are spatially associated with each other: are they close to each other? Another type of related information refers to the definition of their closest common ancestral node (e.g., state, county, or country).

Consider an example borrowed from Batista et al. [5], where a document (D) is geoparsed, yielding the result: Lisboa and Santa Catarina. Then, the first step of geocoding, checking the geo-ontology, yields these results (also highlighted in a partial hierarchical geographic concepts depicted in Figure 2.10):

(i) Lisboa is a municipality;

(ii) Lisboa is somewhere in the municipality of Monção;

(iii) Santa Catarina is a civil parish in the Lisboa municipality;

(iv) Santa Catarina is a street in the Porto municipality;

(v) Santa Catarina is a state in Brazil.

Note that the version of Santa Catarina that appears as a state in Brazil (v) was ruled out, for example, because there were no exact matches for both Lisboa and Santa Catarina.

Analyzing the results of the example above, the closest spatial relation is (i) and (iii) as depicted in Figure 2.10. In fact, there is a direct relation between (iii) and (i). Hence, the geocoding can result in associating the footprint of (iii) to that document, if the aim is to capture the most specific scope.

However, sometimes in a digital library (DL), one might want to associate more than one footprint, in order to represent geographic concepts, for example associating it to the footprint of (i) or even broader scope like Portugal; this would ensure that the various possible scopes of a document are captured (geographic signature) [5].

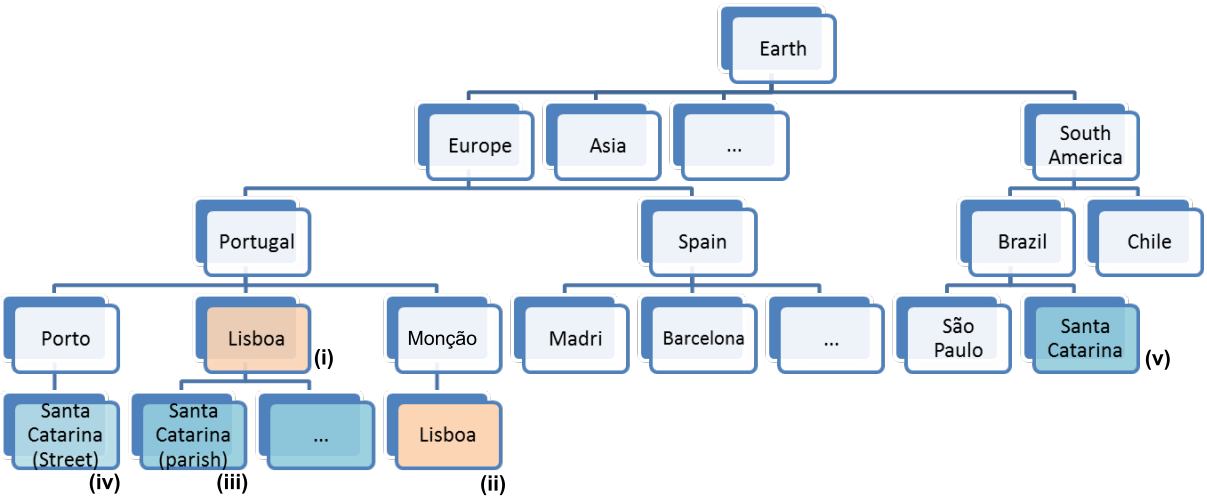


Figure 2.10: A partial hierarchical geographic concepts: Lisboa and Santa Catarina highlighted.

Research Challenges

Research opportunities are related to challenges in the presentation, processing, and data layers as shown in Figure 2.7. Regarding the presentation layer, key concerns relate to how humans can express their information needs through queries, and how they can browse through the results returned by a GIR system. In the processing layer, important challenges are related to the identification and elimination of place name ambiguity, and the design of effective (as well as efficient) algorithms for search, result classification, and ranking. Finally, in the data layer, considering the Web as a huge data repository, the difficulty is to deal with inconsistent and unstructured data to identify and geocode the documents that are found.

Presentation Layer Early computer interfaces forced users to formulate structured queries, similar to what can be supported by a typical database query language (e.g., SQL). However, the majority of users lack knowledge and skill regarding proper use of such structured languages. As a result they do not completely express their needs, and the retrieved information does not fulfill their expectations. Considering also that users have to express spatial notions in words, more complexity and indirection are added to this problem. Often, system results do not fulfill user expectations. All this leads to the

question: Does a query need to be expressed only by words/terms?

The difficulty in designing an interface where users can express themselves informally is related to problems that natural language processing researchers have been tackling for years: ambiguity, imprecision, and human language context dependency. In addition, the imprecision and temporal dependencies attached to the spatial data can make designing a good interface an even more challenging task.

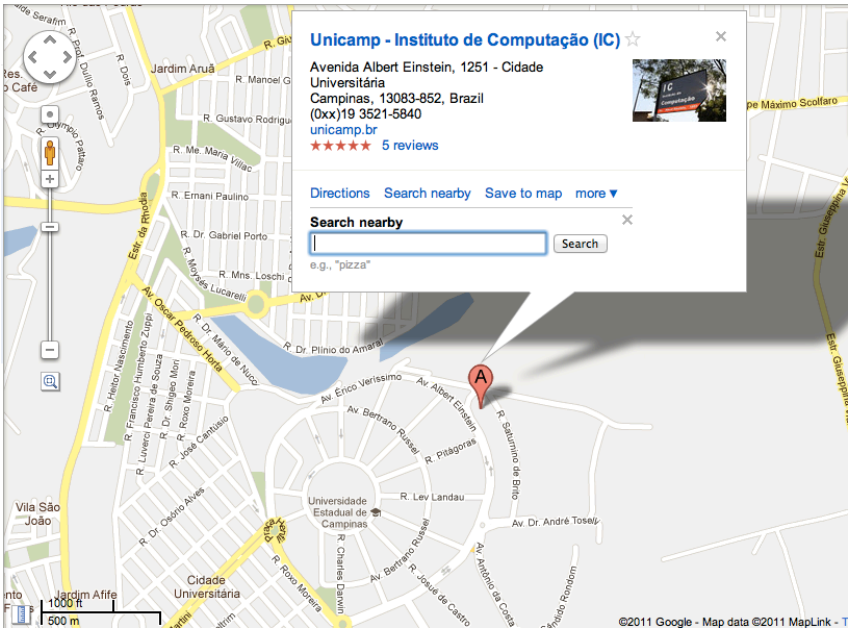


Figure 2.11: Example from Google Maps with a point of interest (POI) selected and search for something nearby enabled.

Consider, as an example of GIR results presentation, the execution of local searches such as those found in Google Places Search. In this case, a set of geocoded pages is retrieved by keyword-based search and they are pinpointed on Google Maps, yielding a geographic view and distribution of the results over a space (Figure 2.12). Moreover, when a point on the map is selected, users are allowed to specify what they want to "Search nearby." One strategy for a GIR query input interface could be using this kind of map-based interface, where it is easy to aggregate queries that involve spatial relationships (Figure 2.11).

There are still many challenges in the presentation layer relating to how geographical results are presented, how users can indicate which results are really relevant (so the system learns how to refine its searching), and how interactions can be improved using visual query interfaces [65, 111] such as those inspired by works developed for GIS: Spatial-Query-by-Sketch [7], and visual formulation of queries based on sets of icons representing

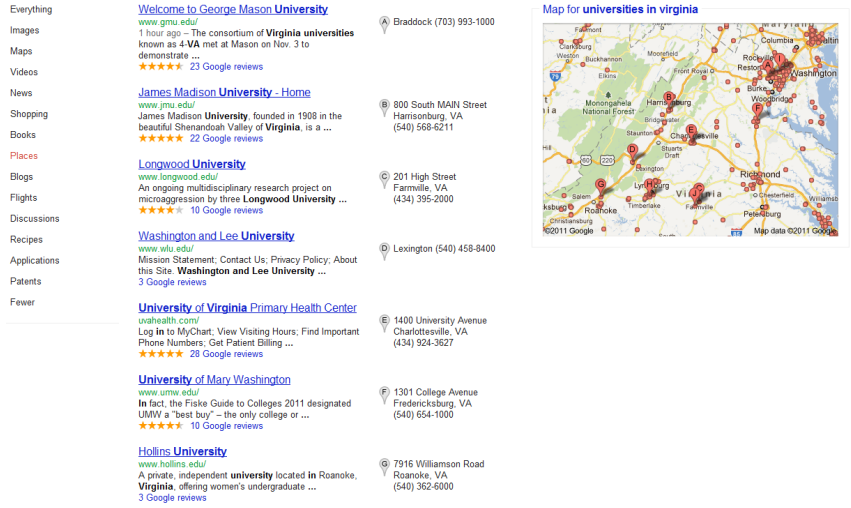


Figure 2.12: Example of results returned by Google Place search.

geographic features and relationships that are combined to build a geographic query [36, 45].

Processing Layer Challenges in the processing layer include identification and elimination of place name ambiguity (e.g., when a common name is used for a variety of places and objects). In this case, the system presents alternative choices (similar names) to the user. According to user feedback, a new query is sent to the system [50, 128].

Thus, supposing users are offered an interface where they can express their queries through semi-structured or natural language, it will be challenging to identify [123], extract [5], and manipulate references to places as well as intended geographical relationships between them [19, 114], and deal with those derived imprecise and ambiguous references [14, 43, 76, 97, 99, 117].

Even if a geographic knowledge base is properly assembled and geocoded, in the light of the huge amount of data spread over the Web, two challenges stand out: efficiently processing queries in a geographical Web search engine [20], and designing effective algorithms to predict if a document is relevant [34, 91, 131].

Moreover, more study is required on users' needs for geographical information, trying to understand their search behavior. For example, by analyzing search logs, Henrich and Luedecke [51] show that searches for geographic information in the USA are often about places to stay and visit, and users intend to buy or rent something from there besides learning how to reach it. In another investigation, the reason why users rewrite Web queries was studied. This aimed to identify users' preferences for physical distances between places searched, as well as the location from which a query was sent [56]. Finally,

the most used geo-words and how they are reused by users was studied by Sanderson and Han [114].

Data Layer Considering the Internet itself as a big data repository, it is challenging to create and index automatically a geographic knowledge base from what is available on the Web [54, 110]. That involves dealing with inconsistent data and also demands to identify and geocode data found in webpages [1, 2, 8, 10, 18].

The data layer includes collections of target documents, as well as supporting structures such as ontologies, thesauri, and geospatial databases, to assist with spatial operations and queries (as discussed earlier). Further research is needed regarding how to build and use such structures and repositories.

In the case of this work, we are tackling a problem in the processing layer. More specifically, we address the problem of geocoding digital objects. This work considers it as a single process that integrates the activity of recognizing place in digital objects (geoparsing) with that of associating them with some place on Earth or with some specific coordinates (latitude and longitude). Moreover, we are interested in geocoding using textual and visual information enclosed in digital objects. Next, we provide an overview regarding the multimedia retrieval of geographic information.

2.1.4 Multimedia Retrieval of Geographic Information

The discussion above of GIR focused on text geocoding and geoparsing. We also argued that only after documents are geoparsed and geocoded can they be placed on a map or queried by geographic location. However, in DLs digital objects go beyond text documents, e.g., they include images and videos.

It is increasing the number of devices connected with GPS and camera, such as smartphones, that embed location data in picture and video metadata, along with other data such as date, time, and camera details. Therefore, it is useful to combine CBIR (content-based image retrieval), multimedia, and GIR techniques in DLs.

Geotagging photos and videos is possible not only when you take or record them from a device with GPS, but it also is enabled by applications and services such as Flickr[3] and Panoramio.[4] In addition to supporting annotation, they allow users to organize and manually assign locations, using a map interface or geographically relevant keywords [89]. Accordingly, the amount of geotagged photos and videos is growing rapidly. For example, in Flickr, there were about 4.7 million geotagged items in 2010 [89], but this number increased to more than 165 million geotagged items by November 2, 2011.

[3] `http://www.flickr.com/` *(as of Dec. 2013).*

[4] `http://www.panoramio.com/` *(as of Dec. 2013).*

Geocoding approaches based on visual clues are proposed in the context of landmark recognition, as well as for non-landmark images [89]. Usually, those approaches are modeled as image classification or content-based image retrieval (CBIR) problems. Those approaches often take advantage of a huge collection of geotagged images that is used as a knowledge base [49]. In [89], the multimedia retrieval for geographic information is used in the following context:

Semantic multimedia understanding encompasses social and cultural semantics, as well as annotation, organization, and retrieval of events, scenes, or objects. For example, white colors associated with a photo of somewhere in the NE of USA during winter indicates snow. Similarly, if a photo depicts people cheering and their location is related to a baseball field, then it may indicate a photo of a baseball game;

Geolocation and landmark recognition aim to determine the location of an image, video, or series of images. In this case, collections of geotagged images are used as training and matching data to help predict the location of unknown images. Landmark images recognition can be seen as detection of somewhat unique objects in unknown images, which are similar to images in a collection of geotagged images. Here matched images' geolocation will aid in the prediction of the location of a given unknown image;

Media visualization can aid the use of collections and landmarks, camera viewing directions, travel trajectories and routes, and photos in large collections (that can be browsed for tourism in 3D fashion);

Recommendation for location-based services or products can help with planning vacations and identifying attractions based on users' locations and interests. This category of applications can be divided further into: real-time recommendation, recommendation inference via geotagged images (considering spatial and temporal patterns), travelogues, and GPS trajectories;

Social network applications: Luo et al. [89] cite works that use tweets or Flickr uploads to discover time and location information related to an event. Users are seen as social sensors; their reports can document the spread of the consequences of an event (such as a flu epidemic or the movement of a typhoon). Therefore, it is important to predict the location of Flickr users. One strategy is based on their social connections' public locations, since users tend to communicate more with closer friends;

Mapping applications can use geocoded photos to produce different kinds of maps, for example, on land use (park, green area, under/super developed area).

As explained above, landmark image recognition is based on detecting unique objects in images and matching them against a knowledge base (collection of geotagged images). This is called landmark recognition with feature point matching, as interest points from a test image are matched to interest points in one or more training set images [89]. However, interest point matching in urban areas is difficult, since some structures (e.g., windows) may repeat frequently. Example of this kind of approach is described in [107], which performs matching of local descriptors to find similar regions within images of a dataset of buildings. Therefore, although they are not explicitly geocoding images, their approaches could be used for that purpose. The approach presented in [107] works with buildings from the University of Oxford.[5] After describing images with a scheme based on a visual vocabulary (quantized local features), a matching strategy is performed between a given query image and images from the dataset. Their approach is robust to deal with changes in illumination, viewpoint, scale, and rotation.

In non-landmark location recognition, image exact match on a training dataset may not occur or may not be reliable. For example, Hays and Efros [49] find a probability distribution of images over the globe and base their strategy on that information, as well as on a dataset of over 6 million geotagged images (their knowledge base) from all over the world. Unknown images are described by selected image descriptors (e.g., color histograms, GIST) and compared to the big knowledge base. The top k most similar returned geotagged images are used to estimate the location of a given unknown image. Although this strategy will not be precise most of the time in finding an exact location, it will indicate roughly where an image was captured. For 16% of the time, their method correctly predicted an image location within 200 km. Extensions of this approach rely solely on the text tags associated with the images [70, 119], or apply Hays and Efros' method to the visual content of images and to their associated user tags [44].

Gallagher et al. [44], besides using a collection of over a million geotagged photographs, also built location probability maps of user tags over the globe to study the picture-taking and tagging behaviors of thousands of users. Applying the local tag probability maps and image matching of Hays and Efros [49], Gallagher et al. showed that their method yielded improvements over pure visual content-based methods. Kalantidis et al. [57] propose geotagging non-landmark images using a big geotagged and clustered dataset as knowledge base.

Similar strategies have been employed in Placing Task at MediaEval,[6] a benchmarking initiative to evaluate a "new algorithm for multimedia access and retrieval", which is a

[5] `http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/` *(as of Dec. 2013).*

[6] `http://www.multimediaeval.org/` *(as of Dec. 2013).*

spin-off of VideoCLEF. We will detail the Placing Task in Section 2.2.

2.1.5 Geographic Information and Digital Libraries

The first large Digital Library (DL) project interested in explicitly using geographic information was ADEPT – Alexandria Digital Earth ProtoType, focused on geocoding DL objects by taking into account textual metadata. It came from the Alexandria Digital Library (ADL), which is a project led by University of California (Santa Barbara, USA) from 1995 to 2004. It is a distributed digital library comprising of a collection of geo-referenced material that could be searched [40, 52]. Its search was focused on its digital library contents. The spatial operators supported in searching this distributed digital library are: "contains item area," "overlaps," "encompasses/contained by," and "exclude/outside."

Many DL initiatives may take advantage of existing geocoding methods. One example is the CTRnet DL [129], as presented in Section 1. As we discussed, in order to provide a map-base browsing or geographical searching services in a DL, its collections must be geocoded. In Section 4.2, a case study that builds the basis to geocode VT's photos [86] will be presented.

2.2 Multimodal Video Geocoding Task

The most common solutions for geocoding multimedia material rely on textual information [72, 89]. Recently, however, more attention has been given to methods that use image/video content in the geocoding process.

Research on video geocoding has been done for the Placing Task from 2010 to 2012. This task was launched in 2010, along with other tasks [72] at MediaEval – a benchmarking initiative to evaluate new algorithms for multimedia access and retrieval (a spin-off of VideoCLEF). The Placing Task aims to automatically assign latitude and longitude coordinates to each of the provided test videos.

Participants in the Placing Task at MedialEval were allowed to use image/video metadata, audio and visual features, as well as external resources, depending on the run submitted. There was one mandatory run in which only visual features could be used to accomplish the task. The organizers of this task released a set of geotagged Flickr videos and images along with their metadata, such as title, tags, and descriptions provided by the owner of that resource, comments of her/his friends, users' contact lists, and other uploaded resources in Flickr. Evaluation was based on the distance to the ground truth geographic coordinate point, in a series of widening circles: 1 km, 10 km, 100 km, 1,000 km, and 10,000 km. Thus, an estimated location is counted as correct at a particular quality level if it lies within a given circle radius.

The approaches for video geocoding submitted to the Placing Task at MediaEval 2010, 2011, and 2012 can be basically divided into methods based on textual information and those based on visual information. In this work, we are interested primarily in methods for combining different modalities of video data to improve the video geocoding results. Therefore we will tend to focus on the multimodal methods or those that use only visual features for the task.

Placing Task 2010 and 2011

In 2010, the Placing Task data set was divided into 5091 videos for training (with the same additional Flickr photos) and 5125 videos for testing.

There were three main approaches [72]: (a) geoparsing and geocoding texts extracted from metadata assisted by a gazetteer such as GeoNames;[7] (b) propagation of the georeference of a similar video in the development database to the test video; and (c) dividing the training set into geographical regions determined by clustering or a fixed-size grid and later employing a model to assign items to each group. The model estimation was based on metadata text data and visual clues. The best result in 2010 for this task was accomplished by VanLaere et al. [70] by only using metadata for images and videos, combining approaches (b) and (c): first a language model identified the most likely area of the video and then the most similar resources from the training set gave the predicted coordinates.

The only group in 2010 that made use of visual features besides the textual data was Kelm et al. [62]. They reported that combining visual and textual results can yield better results than just relying on one of the modalities of information (just text or visual content).

Kelm et al. also presented a hierarchical approach to geocode videos automatically based on both textual and visual information [61]. The proposed method can be divided into the following steps: (1) geographic boundaries are extracted based on Natural Language Processing (NLP) for toponym recognition, and are filtered by Geonames – and Wikipedia-based filters; (2) a textual region model based on a document classification method, which selects regions with higher probability of being assigned, is employed; (3) a visual model based on the similarity of all frames of a video with regard to training set (videos and photos) mean feature vectors of regions is used. The results are then combined based on their rank sum and, finally, the most similar videos from training data contained in selected regions are determined and their coordinates (latitude, longitude) are assigned to the test video. In summary, a geographical boundary extraction reduces the number of possible regions in a first stage. Then the textual model returns the log-likelihoods of the remaining regions based on the tags of each test video. Next, the visual

[7] `http://www.geonames.org` *(as of Dec. 2013).*

model returns the similarities considering the feature vectors of the region model and the test video. Their approach is based on different and well-defined stages with fusion done on rank level using the rank sum algorithm.

In Placing Task 2011, its data release included 10,216 geotagged videos, along with their extracted keyframes and corresponding pre-extracted low-level visual features, and metadata for 3,185,258 CC-licensed "Flickr photos uniformly sampled from all parts of the world" [113]. Test data comprised of 5,347 videos with its related metadata (without latitude and longitude information).

Although a minimum of one run that uses only audio/visual features was required, most of the participants focused on modeling and solving the problem based on text metadata associated with available videos.

In 2011, six groups submitted their results, but only four of them submitted for a run in which only visual features are used to predict the location of test videos: ICSI team [22], WISTUD team [47], UGENT team [70], and UNICAMP team [77]. However, most of them considered visual features as a backup predicting approach for the cases in which no tags or textual description associated with a test video are available. Text-based video geocoding still yields better results than the visual-based ones, being UGENT the best results when considering the run in which they were allowed to use additional crawled data.

Choi et al. [22] (ICSI team) used the top-three results of searches based on textual metadata as anchor points for an 1-NN search using visual features match (GIST). Each test video (its temporal mid-point frame) is compared to the whole development set (photos and video frames) that is within 1 km radius from those anchor point. They also considered acoustic clues in the video geocoding process when matches of textual- and visual-based results were too low.

Using the 2011 database, Kelm et al. [64] extended their previous work [61] introducing a spatial segmentation at different hierarchy levels with a probabilistic model to determine the most likely location at these levels. The world map was iteratively divided into segments of different sizes and those spatial segments for each level were used as classes for their probabilistic model. They used additional external resources like GeoNames and Wikipedia for toponym detection when generating hierarchical segments (e.g., national borders detection). They combined modality in a sequential mode: first used text for geo-predicting, then in case of absence of metadata the visual approach is applied.

Placing Task 2012

In 2012, the best results were accomplished by CEALIST group [109] using textual data and additional external data. Their approach combines a language model that divides the Earth into cells of approximately 1 km^2; and a user model based on tagging probability,

which exploits users' past geotagging behavior. To construct the user model, this team downloaded 3,000 geotagged metadata per user for training purposes. The visual-only submission of the CEALIST team is based on the bag-of-words model with the SURF descriptor. A bag is associated with video frames and are later used to execute 50-NN video searches with the aim of performing spatial clustering within 5 km [109].

The IRISA team [121, 122] approach is based on tag analysis with a fallback system that relies on user information (upload history, social info like friend and his current/prior location, home town). For the run based on visual properties, the team used their proposed descriptors, which are based on SIFT and VLAD. Those descriptors are used to index training videos using product quantization. The final step of the proposed geocoding approach relies on performing a NN-search on the created index aiming at generating a list of candidate videos. That list is then used to define a list of coordinates, whose medoid lat/long is assigned to the test video [121]. They used the visual content as one of the last resources to be used to geocode a video in their sequential pipeline.

Extending their approach proposed in the previous year, the Ghent and Cardiff team (Ghent) [71] relied on clustering tags found in photos of the training set. Additional information was crawled and used in the geocoding process. Later, test videos are classified into the most probable cluster according to χ^2 feature selection. As 2012's test dataset had a shortage of tags, they used other information (title, description) to treat as tags when they are not found. They used as fall back system the default location (either user home location or center of London). Their visual-only solution relied on extracting SIFT features from photos of both training and test set. SIFT feature vectors associated with frames are then compared to find the most similar training photos. The results where textual and visual features were used to geocode did not improve the results of their approach that relies on videos' metadata.

The ICSI team [21] proposed an approach based on a graph model created by using textual tags to infer the location of a given test video. For 1 km precision level, the proposed graph model was not able to outperform the results obtained by their previous year's approach [22] combined with gazetteers. For other precision levels, however, its graph model yields better results. The visual-only submission of this team was based on GIST features.

The TUD team [88] presented a exploratory study only using visual features associated with regions defined by partitioning the Earth based on different external resources, such as climate and bioma data. Their best results (visual) were those in which the world was divided into regions based on bioma data over which the training photos were distributed and then clustered into subregions. They used the visual features provided by the organizers of the Placing Task 2012 [112].

In 2012, Kelm et al. represented the TUB team [63]. They tackled the geocoding

task as a classification problem that considers different hierarchies or spatial segments as explained in [64] and discussed previously in Section 2.2. Their visual-only approach uses visual features extracted from 3.2 million images and from video keyframes of development/training set. Based on their spatial segmentation in different levels, a k-d tree is built iteratively for distinct image descriptors and segmentation level. The most similar spatial segment is determined by traversing the created k-d tree, using the Euclidean norm [63]. Their best result was accomplished using additional resources.

In our work, we do not use any additional or external resources (e.g., gazetteers, more crawling, etc) and barely use the ≈3.2M Flickr image data set. We apply a late fusion approach to combine video features. A rank aggregation method combines scores generated by various features (from different modalities, e.g., text and visual). Therefore, the features are homogeneously and seamlessly combined, representing an important advantage of our approach. Additionally, other new features can be easily added to the fusion step and this approach opens new research opportunities related to the development and use of rank fusion methods in video geocoding tasks. We will present our approach and analyze its results for this task in Section 4.1.

2.3 Data Fusion

In multimedia retrieval tasks, an accurate information fusion of the different modalities is essential for the system's overall effectiveness performance [68]. The main reasoning behind information fusion systems is based on the conjecture [69] that, by combining features, it is possible to achieve a more precise representation of the data being analyzed.

Given the wide range of applications, information fusion has established itself as an independent research area over the last decades [68]. Most of the approaches fall in three broad categories: early fusion, late fusion, and transmedia fusion. The early fusion approach consists in representing multimedia objects in a single feature space. On the other hand, late fusion and transmedia fusion strategies consider each feature independently. Late fusion techniques usually merge the similarity information encoded by a single modality using aggregation functions. In transmedia approaches, one of the modalities is used to obtain relevant objects and later the retrieval system uses the other available modalities to improve the effectiveness of results [24].

A common approach used for information fusion in multimedia retrieval systems is the application of rank aggregation methods. Rank aggregation methods combine scores/rankings generated by different features (from distinct modalities) to obtain a more accurate one. In many situations, rank aggregation has been seen as a way for obtaining a *consensus* ranking when multiple ranked lists are provided for a set of objects.

Different modalities, or even sets of descriptors of a same modality (e.g., color descrip-

tor), may produce different rankings (or similarity scores). Thus, these distinct views of same data may provide different but complementary information about the multimedia objects. The best combinations occur when all systems being combined have good effectiveness performance, although it is possible to get improvements when only one of the systems is effective. This is supported by the statement that the combinations with the lowest error rate are those whose inputs are independent and non-correlated [27].

More formally, rank aggregation can be seen as the task of finding a permutation that minimizes the Kendall-tau distance to the input rankings. The Kendall-tau distance can be defined as the number of pairwise disagreement about their order in two ranked lists, or "the sum over all input rankings of the number of pairs of elements that are in a different order in the input ranking than in the output ranking. If the input rankings are permutations, this problem is known as the Kemeny rank aggregation problem" [116]. Rank aggregation methods have been exploited for a large number of multimedia applications, since there has been an explosion of such type of digital content in the last years [24].

The strategies that have been used in rank aggregation consider mainly two information of a item in a ranked list: (i) the scores computed for it and; (ii) the position (or rank) assigned to it. CombSum and CombMNZ algorithms [39], for example, consider the sum of the normalized relevance scores computed by various systems to compute a new relevance score. On the other hand, the Borda count method [25] uses rank information in voting procedures. Rank scores are assigned linearly to documents in ranked lists according to their positions and are summed directly. Although very simple and presenting linear complexity, these approaches have been used as baselines for many works along decades.

Another traditional approach to the analysis of rank aggregation resides in the Condorcet criterion. The Condorcet voting algorithm defines that the winner of the election is the candidate that beats or ties with every other candidate in pairwise comparisons. In other words, given a distance between two ranked lists as the number of pairs whose elements are ranked reversely, then the Condorcet's result is the one that minimizes the total distance [30].

Additional common approach is based on Markov Chain where items are represented in the various lists as nodes in a graph. The transition probabilities from node to node is defined by the relative rankings of the items in the various lists. The aggregate rankings are computed by a stationary distribution on the Markov Chain, by determining which nodes would be visited more often in a random walk on the graph [118].

Taking as a starting point the traditional initial methods, many variations have been proposed and the rank aggregation approaches have remained in constant evolution. Although still conserving the unsupervised strategy, initial rank aggregation approaches have evolved to more sophisticated algorithms [30, 67, 93]. At the same time, however,

even simple new approaches have been proposed with good effectiveness results [26]. Reciprocal Rank Fusion (RRF) [26], for example, is a simple method for combining the document rankings from multiple IR systems. RRF sorts the documents according to a naïve scoring formula and the reasoning behind it is that while highly ranked documents are more important, the importance of lower-ranked ones does not vanish.

In this work, we use an unsupervised score-based rank aggregation approach for combining ranked lists defined by features of different modalities to improve geocoding result (Section 4.1.1). It is inspired by works that successfully combined textual and visual evidences to improve multimedia retrieval [24, 133]. To the best of our knowledge, the use of rank aggregation methods in video/image geocoding tasks has not been investigated in the literature yet.

Chapter 3

A Rank Aggregation Framework for Multimodal Geocoding

This chapter presents the proposed multimodal framework for digital object geocoding in Section 3.1, and a new evaluation measure for assessing the quality of results provided by geocoding approaches in Section 3.2.

3.1 Proposed Framework for Multimodal Geocoding

Section 3.1.1 formalizes the geocoding process, while Section 3.1.2 presents the architecture which has been implemented to validate the proposed framework, followed by implementation aspects of the framework (Section 3.1.3). For the sake of improving our descriptions, we assume the task of geocoding video collections, without loss of generality. Note, however, that any kind of digital objects could be geocoded using the proposed framework.

In the following subsections, we use as an example a collection whose videos are associated with textual descriptions, along with their visual features. Another assumption is that two collections are available, following the denominations used in the Placing Task of MediaEval: the development set – that might be referred to as training set – whose purpose is to work as a knowledge base that contains ground-truth latitude and longitude coordinates that can be used to guide the geocoding process; and the test set, on which the algorithms are evaluated.

3.1.1 Formalization

Let $\mathcal{C}_{dev}$=$\{v_1, v_2, \ldots, v_{|\mathcal{C}_{dev}|}\}$ be a video collection named *development set*, such that each video $v_i \in \mathcal{C}_{dev}$ has its location, (x_{v_i}, y_{v_i}), defined. Let $\mathcal{C}_{test}$=$\{v_1, v_2, \ldots, v_{|\mathcal{C}_{test}|}\}$ be a video

collection named *test set*, such that the location (x_{v_q}, y_{v_q}) of $v_q \in \mathcal{C}_{test}$ is *unknown*.

The objective of the geocoding process is to assign a proper location to videos $v_q \in \mathcal{C}_{test}$ given the *known* locations available in the development set, i.e., the development set is used as a knowledge base. Our solution to this problem exploits a multimodal video retrieval paradigm in which the location of a test video is determined according to its similarity distance to videos in the development set.

Let $\mathcal{D} = \{D_1, D_2, \ldots, D_{|\mathcal{D}|}\}$ be a set of video descriptors, such that each video descriptor $D_k \in \mathcal{D}$ defines a distance function $\rho : \mathcal{C}_{test} \times \mathcal{C}_{dev} \rightarrow \mathbb{R}$, where $\mathbb{R}$ denotes real numbers. Consider $\rho(x, y) \geq 0$ for all (x, y) and $\rho(x, y) = 0$, if $x = y$. The distance $\rho(v_q, v_i)$ among all videos $v_q \in \mathcal{C}_{test}$, $v_i \in \mathcal{C}_{dev}$ can be computed to obtain a $|\mathcal{C}_{test}| \times |\mathcal{C}_{dev}|$ distance matrix A.

The proposed framework is multimodal if we assume that video descriptors used in $\mathcal{D}$ define distance functions that exploit different modalities (e.g., visual properties, textual descriptions). Examples of various video descriptors are presented in Section 4.1.1.

Given a query video $v_q \in \mathcal{C}_{test}$, we can compute a ranked list R_q in response to the query by taking into account the distance matrix A. The ranked list R_q=$\{v_1, v_2, \ldots, v_{|\mathcal{C}_{dev}|}\}$ can be defined as a permutation of the collection $\mathcal{C}_{dev}$, such that, if v_i is ranked at lower positions than v_j, i.e., v_i is ranked before v_j, then $\rho(v_q, v_i) < \rho(v_q, v_j)$. In this way, videos of the development set are ranked according to their similarity distance to the query video v_q. Note that the proposed formalism can be easily extended to deal with actual similarity scores as they can be defined in terms of distance functions.

We can also take each video descriptor $D_k \in \mathcal{D}$, in order to obtain a set $\mathcal{R}_{v_q} = \{R_1, R_2, \ldots, R_{|\mathcal{D}|}\}$ of ranked lists for the query video v_q.

The geocoding function $\mathcal{G} : \mathcal{R} \rightarrow \mathbb{R}^2$ is used to define the location of a query video v_q, given its ranked lists $\mathcal{R}_{v_q}$:

$$(x_{v_q}, y_{v_q}) = \mathcal{G}(\mathcal{R}_{v_q}). \tag{3.1}$$

The implementation of $\mathcal{G}$ requires the use of an appropriate rank aggregation method to combine the ranked lists defined in $\mathcal{R}_{v_q}$, as well as an strategy to define a location given the final ranked list. The rank aggregation methods evaluated in our experiments are presented in Section 4.1.1. The assigned location, in turn, is defined, in our current implementation, in terms of the top-ranked video $v_i \in \mathcal{C}_{dev}$ in the final ranked list.

3.1.2 Framework Architecture

The proposed architecture for video multimodal geocoding combines the video visual and textual descriptions, defined in terms of descriptors. It is composed of the following modules (Figure 3.1):

1. text-based geocoding: it is responsible for all text processing and may use GIR geocoding techniques to predict a location based on the available textual metadata;

2. content-based geocoding: this module predicts a location based on the visual similarity of the test images/videos with regard to the knowledge database (available training dataset);

3. geo-semantic: this module refers to the use of ontologies, thesauri, or gazetteers to improve the identification of geographic references (e.g., inside text-based geocoding); and

4. data fusion: this module combines the geocoding results generated by the previous modules and computes the final result of the geocoding. The idea is to rely on both textual and visual descriptions whenever possible.

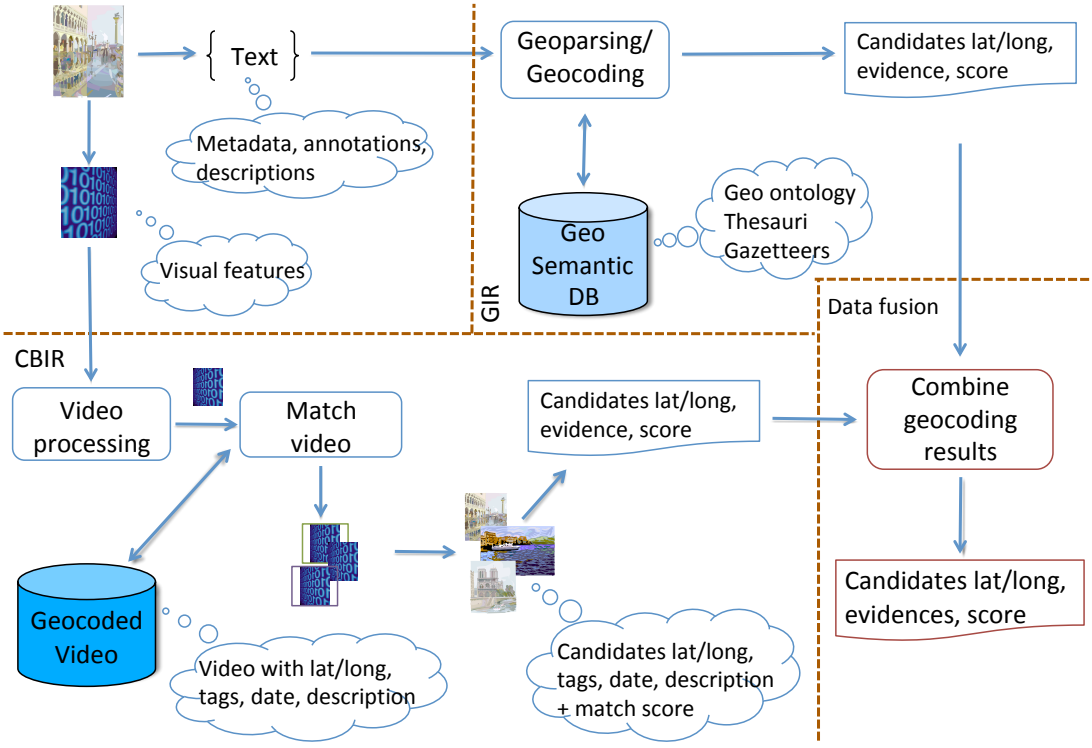


Figure 3.1: Proposed architecture for video multimodal geocoding.

The final result is a combination of the results from each modality, treated by a data fusion module which deals with the geocoding results of textual (metadata) and visual (frames) parts of a video. The modules of this architecture can be developed and evolved independently and later their individual results can be combined by the fusion module.

The idea is that the better the results provided by the individual module, the better should be the final combined results.

Note that the databases represented in the figure are the knowledge base. The one in the content-based geocoding module is derived from a development/training set or from somewhere else. Thus, it is another possible point of customization for this architecture.

In the following section, we present how each module of the proposed architecture has been implemented. The current implementation does not consider the use of ontologies, thesauri, or gazetteers. That is left for future work.

3.1.3 Implementation Aspects

In this work, the modules of the proposed architecture were implemented based on existing content-based and textual information retrieval methods. In fact, the architecture can handle as many different modalities as desired. The final result is a combination of the results from each modality, which are treated by a data fusion module that takes advantage of rank aggregation methods.

Algorithm 1: Multimodal geocoding framework

Input: $\mathcal{C}_{dev} = \{v1, v2, ..., v|\mathcal{C}_{dev}|\}$: video collection in development set,
$\mathcal{C}_{test} = \{v1, v2, ..., v|\mathcal{C}_{test}|\}$: video collection to be geocoded in test set, and
$\mathcal{D} = \{D_1, D_2, \ldots, D_{|\mathcal{D}|}\}$: a set of video descriptors.

Output: *GeocodingResult* as a list of (x_{v_q}, y_{v_q}) for $v_q \in \mathcal{C}_{test}$.

```
1 begin
2     Initialize R;
3     R = ProduceRankedList(C_dev, C_test, D);
4     GeocodingResult = G(R, C_test, C_dev);
5     return GeocodingResult;
6 end
```

A possible deployment of our proposed framework is outlined in Algorithm 1 and the details of the main functions used there are presented in Algorithms 2 and 3. The notation used follows the naming convention defined in Section 3.1.1.

As shown in Algorithm 1, the proposed implementation calls two main components. First, function `ProduceRankedList` (Line 3 of Algorithm 1) is called to process the input videos in order to produce, for each test video, a set of lists, each one ranked according to the similarity of the query video to those in the development set (knowledge base) as detailed in Algorithm 2. Then function $\mathcal{G}$ (called in Line 4) deals with the geocoding itself by combining the resulting set of ranked lists produced previously.

Algorithm 2: `ProduceRankedList`

```
1  ProduceRankedList(C_dev, C_test, D)
   Input: C_dev, C_test, and D as defined for the Input in Algorithm 1.
   Output: a set of ranked lists, one for each test item v_q ∈ C_test and each descriptor
           D_k ∈ D.
2  begin
3      Initialize R;
4      foreach v_q ∈ C_test do
5          Initialize R_{v_q};
6          foreach D_k ∈ D do
               // generate the ranked list for a given v_q according to
                  which feature the descriptor D_k describes (feature vector
                  and distance function)
7              R_{v_q}[D_k] = RankList(v_q, D_k, C_dev);
8          end foreach
9          R[v_q] = R_{v_q};
10     end foreach
11     return R;
12 end
```

As we can observe in Algorithm 2, the `ProduceRankedList` receives as input the knowledge base (development set), the test set to be geocoded, and the set of descriptors to be used. For each item in test set (Line 4 of Algorithm 2), and then for each descriptor (Line 6) a ranked list will be produced. Based on each input descriptor, function `RankList` (Line 7) will calculate the similarity between the test item and all items in the knowledge base, ranking them according to their similarity scores. In the end, this will return (Line 9) a set of ranked lists (one set for each descriptor) for each item in test set. In conclusion, the key for this component is to exploit the list of descriptors, in which the ranked list production will rely on.

In Algorithm 3, we can observe that function $\mathcal{G}$ basically iterates through each test item, taking its corresponding set of ranked lists and calls two functions: `FusionList` (Line 5 of Algorithm 3), which is in charge of combining them, and `geocodeFromRank` (Line 6), which takes as input the resulting list from previous function to estimate the location for the test item handled.

Function $\mathcal{G}$ is defined in Algorithm 3 as we implemented it for the validation of the framework (described in Chapter 4). In this particular case, as detailed in Algorithm 4, we consider that a test item is the same location – provided by the `getLocation` function (Line 4) – of the top ranked (most similar) item from the development set. This video is obtained by calling function `getBestRankID` (Line 3). Note that different strategies

Algorithm 3: Geocoding function $\mathcal{G}$

1 $\mathcal{G}$ ($\mathcal{R}$, $\mathcal{C}_{dev}$, $\mathcal{C}_{test}$)
Input: Set of RankedList $\mathcal{R}$, being one for each descriptor $\in \mathcal{D}$ and $\mathcal{C}_{dev}$ and $\mathcal{C}_{test}$ as defined for the Input of Algorithm 1.
Output: Geocoding result as a list of (x_{v_q}, y_{v_q}) for each $v_q \in \mathcal{C}_{test}$.
2 **begin**
3 Inicialize GeocodingResult;
4 **foreach** $v_q \in \mathcal{C}_{test}$ **do**
5 $NewRankedList_{v_q}$ = `FusionList`($\mathcal{R}[v_q]$);
6 (x_{v_e}, y_{v_e}) = `geocodeFromRank`($NewRankedList_{v_q}, \mathcal{C}_{dev}$);
7 $GeocodingResult[v_q] = (x_{v_e}, y_{v_e})$;
8 **end foreach**
9 **return** *GeocodingResult*;
10 **end**

may be employed to define the location of a test video, given existing ranked lists. In terms of implementation, developers could replace either/both `getBestRankID` or/and `getLocation` functions. For example, instead of taking the best ranked item to estimate the test video coordinates, one could devise other approaches based on the top-k best ranked items.

Another point of modification relies on the use of different implementations of the `FusionList` function (Line 5 in Algorithm 3). In Chapter 4, three different fusion approaches, defined according to the equations described in Section 4.1.1, were tested.

Algorithm 4: Function `geocodeFromRank`

1 `geocodeFromRank`($RankedList_{v_q}, \mathcal{C}_{dev}$)
Input: Ranked list $RankedList_{v_q}$ for a query video v_q and $\mathcal{C}_{dev}$ as defined for the Input in Algorithm 1.
Output: coord=(x_{v_i}, y_{v_i}): predicted coordinate.
2 **begin**
3 v_i=$RankedList_{v_q}$.`getBestRankID` ();
4 coord=$\mathcal{C}_{dev}$.`getLocation` (v_i);
5 **return** *coord*;
6 **end**

3.2 Weighted Average Score (WAS): a Novel Evaluation Measure for Geocoding Tasks

Two evaluation criteria are considered in our experiments. The first one is the most commonly used method to assess the effectiveness of the results submitted to the Placing Task (Section 4.1.2). The second evaluation measure is defined in this section and is one of our contributions in this work. It has never been used to evaluate video geocoding methods before.

Besides the way each approach is evaluated by MediaEval 2012, in this work we propose a new scoring method whose goal is to assess the overall performance of the method based on those geographic distances. The Weighted Average Score (WAS) gives higher weights to the predictions with higher precision.

In other words, instead of a table with accumulative count, we propose a score between 0 and 1 to indicate an overall expected precision level for a geocoding method being evaluated. This proposed score follows the principles of utility theory [33, 37]. According to utility theory, there is a utility function (a user's preference function) that assigns a utility value (the gained value from a user's perspective) for each item. These values vary from item to item. The item can be a book, a product, or a video, as in our case. In general, we assume the utility of a relevant video decreases with its ranking order. More formally, given a utility function $U(x)$, and two positions x_1, x_2, with $x_1 < x_2$, according to this assumption, we expect the following condition to hold: $U(x_1) > U(x_2)$. There are many possible functions that can be used to model this utility function satisfying the order-preserving condition given above.

Given a test set $\mathcal{C}_{test}$, let $d(i)$ be the geographic distance between the predicted location and the ground truth location of the video $i \in \mathcal{C}_{test}$.

The proposed score for the result of a given test query i is defined as follows:

$$score(i) = 1 - \frac{\log(1 + d(i))}{\log(1 + R_{max})}, \tag{3.2}$$

where R_{max} is the maximum distance between any two points on the Earth's surface. The length of the half of Earth's circumference at the Equator is 20,027.5 km, thus we assume $R_{max} = 20,027.5$. The *log* function is used to reduce the impact of different distances observed in the interval $d(i) \in [0, R_{max}]$. Observe that $score(i)$ ranges from 0 to 1, where 1 indicates a perfect estimation ($d(i) = 0$); and 0, an incorrect prediction ($d(i) = R_{max}$). The other *score* values give a sense of how good was the location estimation with regard to the ground truth when it is closer to 0 or to 1. The curve showing the behavior of this function is presented in Figure 3.2.

Let R be a ranked list of $i \in \mathcal{C}_{test}$ ordered by $score(i)$ and let v_i be the item at

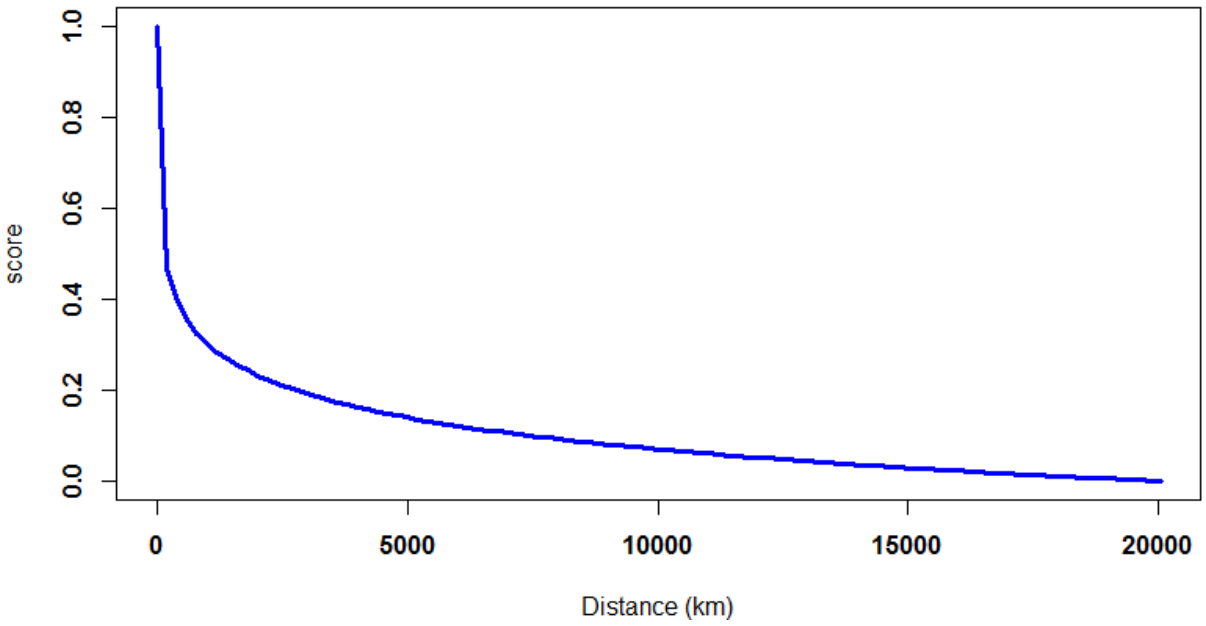


Figure 3.2: The curve of score(i).

the top of this list, which is followed by item v_j. Therefore, they satisfy the condition $score(v_i) \geq score(v_j)$. Note that our scoring formula is based on $d(i)$ rather than the position of the item in a list. So we can say that our utility function can be defined as following: $U(x_i) = score(vid(x_i))$, where function $vid()$ retrieves the item ($v_i \in \mathcal{C}_{test}$) that is ranked at position x_i and that has resulted in $score(v_i)$. In summary, the geocoding result with lower distance, thus higher $score(v_i)$, is indeed highly preferred over the one with higher distance (and lower score) as shown in Figure 3.2.

Finally, let $\mathcal{C}_{test}$ be a test dataset with n videos whose locations need to be predicted. The overall score for the predictions of a method m for $\mathcal{C}_{test}$ is defined as:

$$\text{WAS(m)} = \frac{\sum_{i=1}^{n} score(i)}{n} \tag{3.3}$$

This scoring system helps to tackle some issues that might arise when we look at an accumulative count table for each of the 1, 10, 100, 1 000, and 10 000 km classes (the distance from the estimated geographic coordinate of a video to its ground truth). Consider the examples in Tables 3.1, 3.2, and 3.3 with the hypothetical results for three methods (a, b, and c) experimented on a test set with 100 queries.

The first and last columns are usually reported when the results are shown in Placing Task style. However, to help the understanding, we added two more columns in those tables. The second column gives us the average distance for those test queries in the corresponding precision level. The third column (Average $score(i)$) presents the average

Table 3.1: WAS(a) vs. Accumulative Count.

WAS(a) = 0.625888			
Precision Levels (km)	**Average Distance (km)**	**Average *score(i)***	**Accum. count**
≤ 1	0.50	0.959066	30
≤ 10	5.00	0.819113	50
≤ 100	50.00	0.603063	65
≤ 1000	500.00	0.372403	80
≤ 10000	5000.00	0.140127	100

Table 3.2: WAS(b) vs. Accumulative Count.

WAS(b) = 0.640495			
Precision Levels (km)	**Average Distance (km)**	**Average *score(i)***	**Accum. count**
≤ 1	0.50	0.959066	25
≤ 10	5.00	0.819113	60
≤ 100	50.00	0.603063	65
≤ 1000	500.00	0.372403	80
≤ 10000	5000.00	0.140127	100

Table 3.3: WAS(c) vs. Accumulative Count.

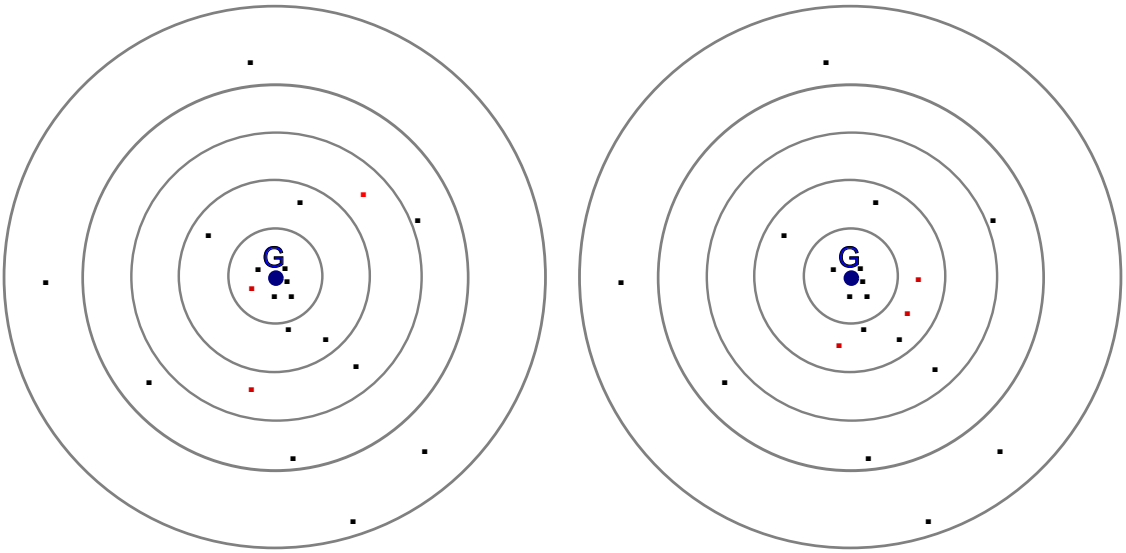
WAS(c) = 0.659540			
Precision Levels (km)	**Average Distance (km)**	**Average *score(i)***	**Accum. count**
≤ 1	0.50	0.959066	25
≤ 10	2.50	0.873527	60
≤ 100	50.00	0.603063	65
≤ 1000	500.00	0.372403	80
≤ 10000	5000.00	0.140127	100

$score(i)$ for the results within that radius/distance.

In order to support the discussions about the examples given by the tables, Figure 3.3 and Figure 3.4 depict some possible distributions of geocoding result (in widening radius) for some hypothetical test queries with the same ground truth (G). These distributions exemplify valid scenarios for the results shown in Tables 3.1, 3.2, and 3.3, besides following the same proportions of points in each precision radius as is exhibited in the tables.

Tables 3.1 and 3.2 show two methods whose results differ in terms of the number of correctly geocoded test queries within 1 km and 10 km radii. If one cares only about results in 1 km, surely one would consider the method a as the best one because of its higher count in that precision level. Nonetheless, for 10 km, the method b is better for the same reason, while for 100 km radius they are both tied with the same amount of geocoded items in that precision level. In this case, the WAS will indicate that the results from method b are better due to: (i) the count difference between them in 1 km is smaller than the 10 km radius, (ii) the disagreement in term of $score(i)$ for items in 1 km or

10 km is small, and (iii) less items in 100 km and more on 10 km radius (as is highlighted in Figure 3.3).

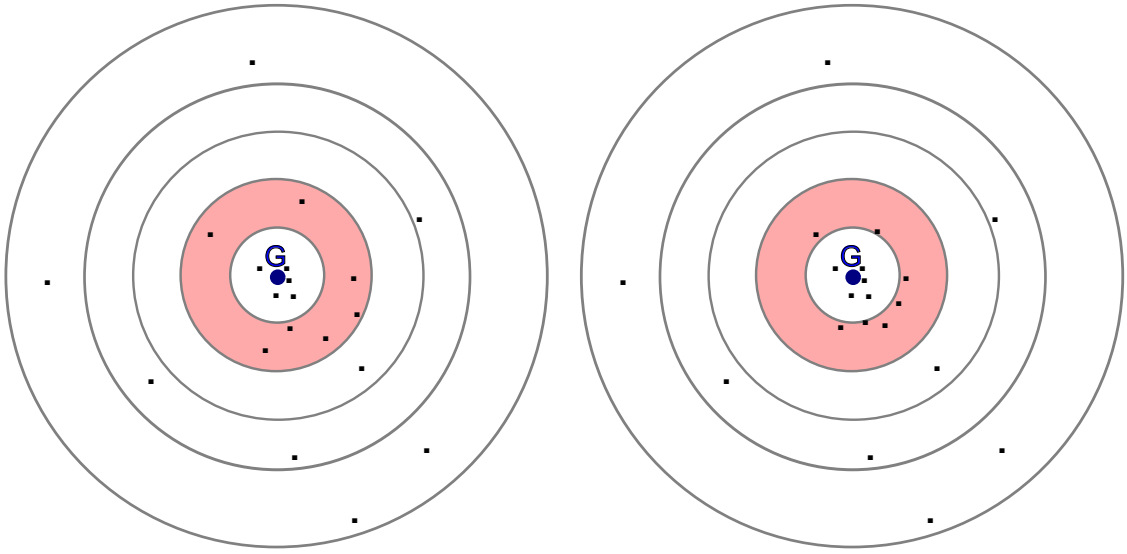


(a) method *a* (b) method *b*

Figure 3.3: Example of geocoding result distribution in various precision radii for methods *a* and *b*.

Tables 3.2 and 3.3, show an example in which the accumulative count in different precision levels is exactly the same for both methods *b* and *c*. The difference will show up when we look into the actual distances between the points that the method predicted and their ground truth. A closer look at the tables, we can notice that at 10 km radius, the average distance for items in that level is 5 km for method *b* and 2.5 km for method *c* (exemplified and highlighted in Figure 3.4). Precisely, this difference will make *WAS(c)* higher than *WAS(b)*. As conclusion, not only does WAS consider these multiple levels of the results, but also the proposed effectiveness measure takes into account every single result of the whole test set to indicate and summarize the level of precision of an evaluated method.

In the experiments described in Chapter 4, we will apply the WAS score to compare and analyze geocoding results using different descriptors.



(a) method b (b) method c

Figure 3.4: Example of geocoding result distribution in various precision radii for methods b and c (spatial distribution difference is highlighted).

Chapter 4

Framework Validation

In this chapter, we present some implementation aspects and discuss results related to the use of the proposed architecture in two geocoding tasks: the Placing Task at MediaEval 2012 and the geocoding of VT Buildings photos.

4.1 Video Geocoding at MediaEval 2012

The aim of conducted experiments is to evaluate the proposed framework for multimodal geocoding in the context of the Placing Task at MediaEval 2012, which is dedicated to the problem of video geocoding. Next section describes this task, as well as the datasets used in the experiments. In the following subsections, we present our strategies to address the proposed task and discuss obtained results.

4.1.1 Architecture Implementation

This section describes the components of the proposed architecture that are considered in our experiments.

Text Retrieval & GIR

For processing available textual information of videos, we propose the use of (1) Geographic Information Retrieval (GIR) techniques to recognize and associate a location with digital objects based on their textual content; and (2) Information Retrieval (IR) classical matching functions to retrieve similar digital objects.

In the context of the Placing Task, we exploit the following metadata associated with videos: title, description, and keywords. In our current implementation, text processing is

based on classical IR text matching using the vector space model and traditional similarity functions [90]: cosine, bag-of-words (normalized documents terms intersection), dice, okapi, and tf-idf-sum.

Let C be a collection with j distinct t terms of index t_j. According to the vector space model, a document d_i is represented as a vector: $d_i = (w_{i1}, w_{i2}, \ldots, w_{it})$, where w_{ij} is the weight of the term t_j in the document d_i. The term weights for a document are often calculated as the $tf \times idf$ value, where tf is term frequency and idf is inverse document frequency of the term in the collection. The idf value is calculated as $log(N/nt)$, where N is the number of documents in the collection and nt is the number of documents that have at least one occurrence of the term t_j.

The textual similarity functions that we used in our framework is the versions cited and implemented by Calumby et al. [16] for cosine, bow, okapi, dice, and tf-idf-sum, which are defined next as presented in [15, p. 4 & p. 37].

$$cosine(d_1, d_2) = \frac{\sum_{i=1}^{t} w_{1i} \times w_{2i}}{\sqrt{\sum_{i=1}^{t} w_{1i}^2 \times \sum_{i=1}^{t} w_{2i}^2}}, \tag{4.1}$$

where w_{ij} is the document as previously defined. Equation 4.1 basically calculates the cosine between the vectors of each document. The closer the cosine is to 1, the more similar the documents are.

$$bow(d_1, d_2) = \frac{|\{d_1\} \bigcap \{d_2\}|}{|d_1 + d_2|}, \tag{4.2}$$

where $\{d_i\}$ is the set of terms that occur in the document d_i. This is a simple measure of the percentage of common words between two documents.

$$dice(c, d) = \frac{2 \times |c \cap d|}{|c| + |d|} \tag{4.3}$$

The dice equation measures the similarity of a document d with regard to a query c based on the number of common terms in relation to the total of terms in both document and query.

$$\begin{aligned} okapi(d_1, d_2) = \sum_{t \in d1 \bigcap d2} & \frac{3 + tf_{d2}}{0.5 + 1.5 \times \frac{size_{d2}}{size_{med}} + tf_{d2}} \\ & \times \log \frac{N - df + 0.5}{df + 0.5} \times tf_{d1}, \end{aligned} \tag{4.4}$$

where tf is the term frequency in the document, df is the term frequency in the collection, N is the number of documents in the collection, $size_{di}$ is the size of document i, and $size_{med}$ is the average document size in the collection.

Given a query c with n terms t_i, the tf-idf-sum is given by the sum of the tf-idf values for each query term in relation to the document d of the collection.

$$tf\text{-}idf\text{-}sum(c,d) = \sum_{i}^{n} tf(t_i,d) \times idf(t_i) \tag{4.5}$$

Visual Information Retrieval

To encode video visual properties, we have used two main approaches. One is based on video frames and does not consider transitions between them, which is called *bag-of-scenes* [105]. The other approach specifically encodes motion information by using the *histogram of motion patterns* [3].

Bag-of-Scenes (BoS)

One of our approaches to encode video visual properties is based on a dictionary of scenes [105]. The main motivation for using the bag-of-scene model is that video frames can be considered as a set of pictures from places. Pictures may contain important information regarding place location. Therefore, if we have a dictionary of pictures from places (scenes), we can associate each video frame with the most similar pictures in the dictionary. The final video representation will then be a place activation vector making it representative for the geocoding task.

An important advantage of the bag-of-scene model is that the dictionary is composed of visual words carrying more semantic information than the traditional dictionaries based on local descriptions. In the dictionary of scenes, each visual word is associated with pictures of a place [105]. A consequence of this property is that the bag-of-scenes feature space has one dimension for each semantic concept, making it easier to detect the presence or absence of the concept in the video feature vector.

The process of creating a dictionary of scenes is similar to the one used to create a dictionary of local descriptions. The main difference is that the feature vectors represent the whole images and not local patches. Practically speaking, instead of quantizing SIFT space, we quantize the bag-of-words space, for example. Thus, each visual word is an image feature vector and not a local patch feature vector.

After creating the dictionary of scenes, the steps to represent a video are the same employed when a dictionary of local descriptions is used to represent an image. In the former, a video is a set of frames. In the latter, an image is a set of local patches. Therefore, we use the popular approaches, like *hard* and *soft* assignment [125], to assign a video frame to the scenes in the dictionary. Next, we apply pooling strategies, like *average* and *max* pooling [12], to summarize the assignments and create the video feature

vector, which is called the *bag-of-scenes*. Thus, comparisons between two bags-of-scenes are performed using the Euclidean distance function.

In [105], the bag-of-scene model is evaluated considering two possibilities to create the dictionary. One uses the video frames of the training set as scenes (BoF) and the other uses an external image dataset (BoS). The results for both representations are very similar. We refer the reader to [105] for details concerning the evaluation of different parameters in the bag-of-scene model, like the dictionary size, the coding, and pooling strategies, as well as the use of different low-level descriptors to represent each video frame.

Histogram of Motion Patterns (HMP)

Besides encoding visual properties using a dictionary of scenes from places of interest, we also adopted a simple and fast algorithm to compare video sequences, described in [3]. It consists of three main steps: (1) partial decoding; (2) feature extraction; and (3) signature generation.

For each frame of an input video, motion features are extracted from the video stream. For that, 2×2 ordinal matrices are obtained by ranking the intensity values of the four luminance (Y) blocks of each macroblock. This strategy is employed for computing both the spatial feature of the 4-blocks of a macroblock and the temporal feature of corresponding blocks in three frames (previous, current, and next). Each possible combination of the ordinal measures is treated as an individual pattern of 16-bits (i.e., 2-bits for each element of the ordinal matrices). Finally, the spatio-temporal pattern of all the macroblocks of the video sequence are accumulated to form a normalized histogram. For a detailed discussion of this procedure, refer to [3].

The comparison of histograms can be performed by any vectorial distance function like Manhattan (L_1) or Euclidean (L_2) distances. In this work, we compare video sequences by using histogram intersection, which is defined as

$$d(\mathcal{H}_{V_1}, \mathcal{H}_{V_2}) = \frac{\sum_i \min(\mathcal{H}^i_{V_1}, \mathcal{H}^i_{V_2})}{\sum_i \mathcal{H}^i_{V_1}},$$

where $\mathcal{H}_{V_1}$ and $\mathcal{H}_{V_2}$ are the histograms extracted from the videos V_1 and V_2, respectively, processing their i video sequences. This function returns a real value ranging from 0, for situations in which those histograms are not similar at all, to 1 when they are identical.

Data Fusion

The data fusion module aims at combining the similarity scores of different modalities, producing a more accurate one. Given a query video (whose location is unknown), it

is compared with all those of the knowledge dataset (training set), considering different features associated with different modalities. Each feature, in turn, produces a different score. The goal of the data fusion module is to combine the scores produced by features of different modalities in order to produce a more effective score. In this work, we evaluated three rank aggregation methods in the video geocoding task.

The first one is based on a multiplication of scores, initially proposed in [103] for multimodal image retrieval. The method was evaluated in several image retrieval tasks related to the combination of image descriptors and combination of visual and textual descriptors. That experimental evaluation considered fifteen visual descriptors (considering shape, color, and texture descriptors) and six textual descriptors with good results.

Let v_q be a query video that is compared to another video v_i in the dataset. Let $sim(v_q, v_i)$ be a function defined in the interval $[0, 1]$ that computes a similarity score between the videos v_q and v_i , where 1 denotes a perfect similarity. Let $\mathcal{S} = \{sim_1, sim_2, \ldots, sim_m\}$ be a set of m similarity functions defined for the different features considered. The new aggregated score sim_a is computed by multiplying individual feature scores as follows:

$$sim_a(v_q, v_i) = \frac{\sqrt[m]{\prod_{k=1}^{m}(sim_k(v_q, v_i) + 1)}}{m} \tag{4.6}$$

By multiplying the different similarity scores, high scores obtained by one feature are propagated to the others, leading to high aggregated values. The reasoning behind the multiplication approach is inspired by the Naïve Bayes classifiers [100, 103, 132]. In a general way, Naïve Bayes classifiers work based on the probability of an instance being of a class, given a set of features and assuming conditional independence among features. In a simplified manner, that classifier assumes that the presence of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Under the independence assumption, the probabilities of each feature being of a given class are multiplied. In this case, as an analogy, the proposed multiplication approach can be seen as the computation of the probability of videos v_q and v_i to be similar, considering independent features.

We also evaluated the traditional Borda [130] approach and the recently proposed Reciprocal Rank Fusion (RRF) [26]. Both methods consider the rank information, i.e., the positions of images in ranked lists produced by different descriptors.

Let $\mathcal{D}$=$\{D_1, D_2, \ldots, D_m\}$ be a set of descriptors and let a v_q be a query video. For each descriptor $D_j \in \mathcal{D}$ we can compute a different ranked list τ_{q,D_j} for the video query v_q. A given video v_i is ranked at different positions (defined by $\tau_{q,D_j}(i)$) according to each descriptor $D_j \in \mathcal{D}$. The objective is to use these different rank data to compute a new distance between video v_q and v_i.

The Borda [130] method considers directly the rank information for computing the new distance $F_{Borda}(q, i)$ between video v_q and v_i. Specifically, the distance is scored by the number of videos that are not ranked higher than it in the different ranked lists [66]. The new distance can be computed as follows:

$$F_{Borda}(q, i) = \sum_{j=0}^{m} \tau_{q,D_j}(i). \tag{4.7}$$

The Reciprocal Rank Fusion also uses the rank information for computing a similarity score between video v_q and v_i. The scores are computed according to a naïve scoring formula:

$$F_{Reciprocal}(q, i) = \sum_{j=0}^{m} \frac{1}{k + \tau_{q,D_j}(i)}, \tag{4.8}$$

where k is a constant. In our implementation we used $k = 60$, as suggested by the original paper.

4.1.2 MediaEval 2012

This section introduces the Placing Task in the MediaEval 2012 initiative.

Datasets

The datasets provided by the MediaEval 2012 organizers for the Placing Task are composed of a development and a test set [112]. The *development* set contains 15,563 videos and 3,185,258 CC-licensed images from Flickr.[1] All of them are accompanied by their latitude and longitude information, as well as title, tags, and descriptions provided by the owner of that resource, comments of her/his friends, users' contact lists and home location, and other uploaded resources on Flickr. Videos are provided with their extracted keyframes and corresponding pre-extracted low-level visual features, and metadata. More than 3 million images available are from all parts of the world. Also, pre-extracted low-level visual features of each image are available. The *test* set comprises 4,182 videos, their keyframes with extracted visual features, and related metadata (without geographic location). The distribution around the world of the videos in the training set and test set are shown in the heat maps in Figures 4.1 and 4.2, respectively. Those heat maps were generated using the tool provided by sethoscope.net[2] and we used its option that employs

[1] `http://www.flickr.com/` *(as of Dec. 2013).*

[2] `http://www.sethoscope.net/heatmap/` *(as of Dec. 2013).*

tiles in toner design, provided by Stamen Design, under CC BY 3.0, and background tiles data from OpenStreetMap,[3] under CC BY SA.

Figure 4.1: Heat map of the distribution of the videos in *training set*.

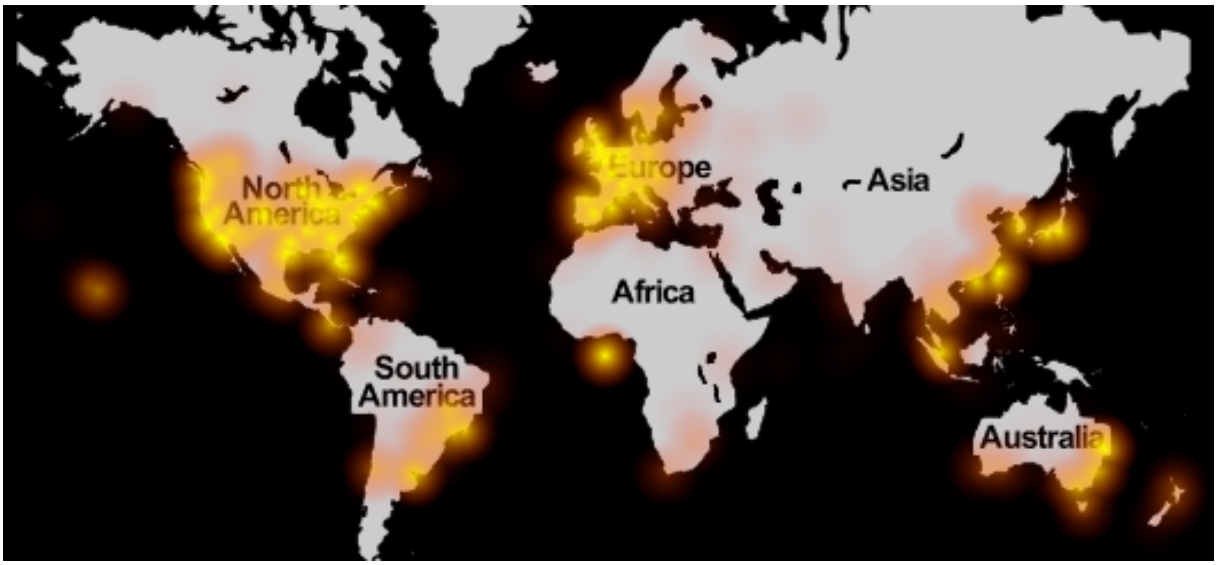


Figure 4.2: Heat map of the distribution of the videos in *test set*.

The keyframes were extracted by the organizers in 4-seconds intervals from videos and saved as individual JPEG-format images. The following visual feature descriptors for

[3] `http://www.openstreetmap.org/` *(as of Dec. 2013)*.

keyframes and photos were extracted and provided: Color and Edge Directivity Descriptor (CEDD), Gabor Texture, Fuzzy Color and Texture Histogram (FCTH), Color Histogram, Scalable Color, Auto Color Correlogram, Tamura Texture, Edge Histogram, and Color Layout. Based on our preliminary experiment, the CEDD descriptor yielded the best results, therefore the visual feature BoS and BoF was based on CEDD descriptor.

Participants in the Placing Task at MediaEval 2012 were allowed to use image/video metadata, audio and visual features, as well as external resources, depending on the run submitted. At least one run should use only audio/visual features.

The experiment reported here concerns the implementation of the proposed architecture by integrating independent work in information retrieval (textual), content-based image and video retrieval (visual), and rank aggregation as summarized in Sections 4.1.1, 4.1.1, and 4.1.1.

Our team used only resources provided by the Placing Task 2012 organizers and we did not make use of any external resources like gazetteers, Wikipedia, or additional crawling. Thus it is fair to compare these results with other teams' equivalent results, which we will highlight later.

In fact, the image collection was only used by the visual feature BoS to sample images for its dictionary of bag-of-scene. Besides that, all the other methods relied only on the 15,563 videos (development set) as their geo-profile database.

Placing Task Evaluation Criteria

According to the evaluation criterion defined in the Placing Task 2012, the effectiveness of a method is based on the great circle distance (Haversine) of the estimated geographic coordinate of a video to its corresponding ground truth location, in a series of widening circles of radius (in km): 1, 10, 100, 1000, and 10000. Thus, an estimated location is counted as correct if it is within a particular circle. In other words, we measure the quality or precision level of correct predictions within a given circle radius.

The results are often reported using a table with an accumulative count of correctly assigned videos at each precision level. This table shows a given method's behavior at different precision levels, for example, in which radius level an evaluated method is able to perform with satisfactory performance. However, when comparing methods, the participants of the Placing Task usually prefer to emphasize the results of smaller circle radii. In that case, we are more interested in determining as accurately as possible the location of a video. More details about Placing Task at MediaEval 2012 are given in the working notes of the organizers [112].

4.1.3 Experimental Setup

Our method to geocode a test query video is composed of three steps: text processing, visual content processing, and data fusion. We used 15,563 videos from the development set (training set) released by organizers of the Placing Task as geo-profiles to which each test video is compared.

In order to assess our proposed framework, we first evaluate our results using only one modality of a video content (textual or visual). In this phase, different (textual and visual) descriptors are used, so the descriptor yielding better results can be used in the information fusion module.

The visual processing module encodes visual content properties of each provided video. Next, the distances between each video in the test set and all videos in the training set are computed. Finally, for each test video, a ranked list of training videos is produced. The text processing module works similarly, except for the feature extraction step, which is based on video textual metadata. In summary, each module produces ranked lists of videos that are then processed by the information fusion module.

In our geocoding scheme, we consider that the query video will receive the lat/long of the top-ranked (most similar) training video.

We also report the results for the development set, that is, we perform experiments considering videos of the development set as query videos. In this case, given that the query video always is the best match to itself (thus it will be the first in this list), we use the second video of available ranked lists to define the final location. We can see it as leave-one-out cross-validation [46, sec. 7.10.1, p. 242], in which each time a different video in the training set is left out and used as a query against the remaining videos in that set.

Regarding the implementation, for textual processing, we set up a Solr[4] server, then we used Python Solr API[5] to index (with stemmer and tokenizer) and generate the corresponding term vectors. Later, they were accessed using a Java program, applying the corresponding version of Apache Lucene Core,[6] to calculate the textual similarities described in Section 4.1.1. The other modules and algorithms were implemented using C, shell scripts, and Python. Finally, the result analyses and evaluations are generated both in R and Python.

[4] `http://lucene.apache.org/solr/` *(as of Dec. 2013).*
[5] `https://github.com/tow/sunburnt` *(as of Dec. 2013).*
[6] `http://lucene.apache.org/core/` *(as of Dec. 2013).*

4.1.4 Results

In this section, we present the results of our experiments. First, we present the results when using a single modality to describe the videos in both development and test sets. These results provide insights about the most suitable descriptors for the geocoding task. Then we perform some correlation analysis on the results for the methods used, showing their potential for combination, bearing in mind that low correlated good results are more likely to produce a good combined result. Finally, we present the results considering the combination of the distinct methods.

Single modality results

We have performed experiments in the development and test sets using each modality (text and visual) in isolation from the other. The objective of this experiment is to determine which descriptor/approach is appropriate for video geocoding.

For textual data, we applied similarity functions as described in Section 4.1.1 over metadata associated with available videos: title, description, and keywords.

Notice that a deeper analysis of the bag-of-scenes (BoS) and the histogram of motion patterns (HMP) approaches are presented in [105] and in [77], respectively. In our experiments, we have used their best parameters.

For the bag-of-scenes method, we performed experiments with dictionaries of 50, 500, and 5000 scenes. Additionally, we considered two different inputs for creating scene dictionaries: the Flickr photo collection and the frames of videos of the development set. We named BoS^{50}_{CEDD}, BoS^{500}_{CEDD}, and BoS^{5000}_{CEDD} for bag-of-scenes with dictionaries based on Flickr photos, and BoF^{50}_{CEDD}, BoF^{500}_{CEDD}, and BoF^{5000}_{CEDD} for those with dictionaries based on frames of videos from development set.

Figure 4.3 and Figure 4.4 show stacked bars associated with evaluated methods considering the development and test sets, respectively. Each rectangle in a stack represents one radii in the set of widening circles (1, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, and 10000 km) traditionally used by the organizers to measure the performance of video geocoding methods in the Placing Task. In those figures, the textual descriptors results are colored in red, while visual results are in green. Darker colors mean smaller radii, therefore, the larger the darker rectangles are, the more precise the evaluated method is. For example, the first rectangle in the bottom of the stack refers to the 1 km radius. The bars related to predictions that are more than 10,000 km from the ground truth location are not shown.

In the development set (Figure 4.3), we can clearly see a better performance for text-based approaches in relation to visual-based approaches. As we can observe, for those methods, more video locations are predicted correctly in more precise (lower) radii. The

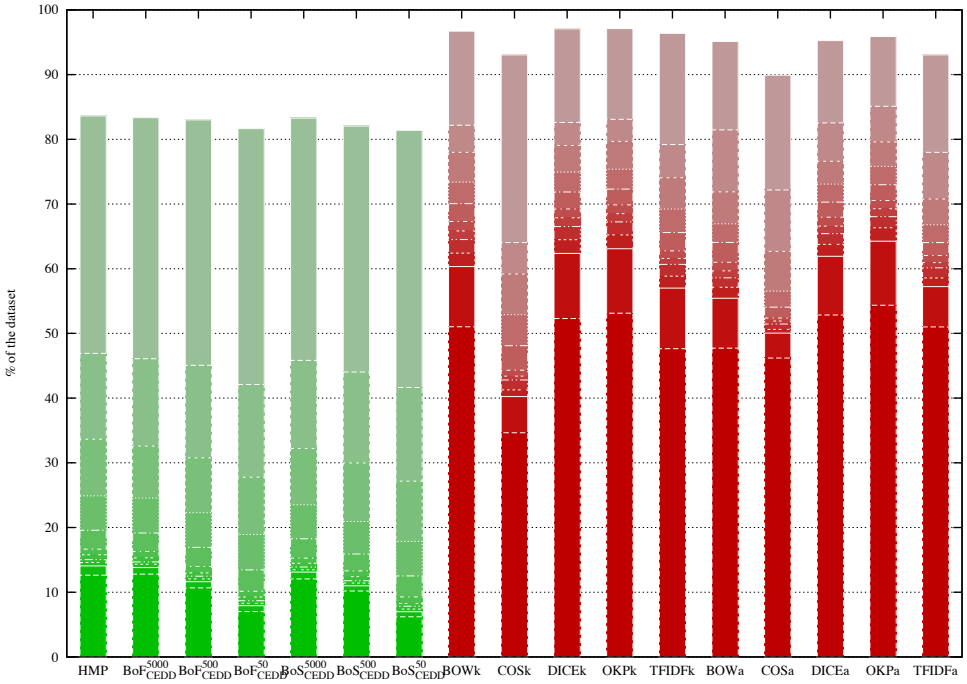


Figure 4.3: Stacked bars showing the isolated performances of each method in the *development* set.

Okapi distance function considering the title, description, and keywords associated with a video (OKPa), or just using keywords (OKPk) yields the best results for 1 km precision, followed by Dice only using keywords (DICEk). Considering only the visual-based approaches, HMP is slightly better than BoF^{5000}_{CEDD}.

In the test set (Figure 4.4), OKPa is again the best method. For visual-based approaches, there are very small differences among the methods, but HMP is still slightly better.

As expected, the results for the test set are worse than those observed for the development set. Most of the text-based approaches are able to geocode into the 1 km widening circle for about 50% of the videos of the development set. For the test set, however, none of them are able to predict very accurately the correct locations. Less than 10% of the videos are geocoded within the 1 km radius (first rectangle of each stack).

Figures 4.5a and 4.5b summarize the performance of evaluated methods, now using our proposed score *WAS(m)* (Equation 3.3) for development and test sets, respectively. As we can observe, the conclusions are similar to those drawn for Figure 4.3 and Figure 4.4: Okapi ($OKPa$ and/or $OKPk$) is the best text-based method followed by Dice ($DICEk$ and/or $DICEa$); while HMP is the best visual-based method followed by BoF^{5000}_{CEDD} and BoS^{5000}_{CEDD}.

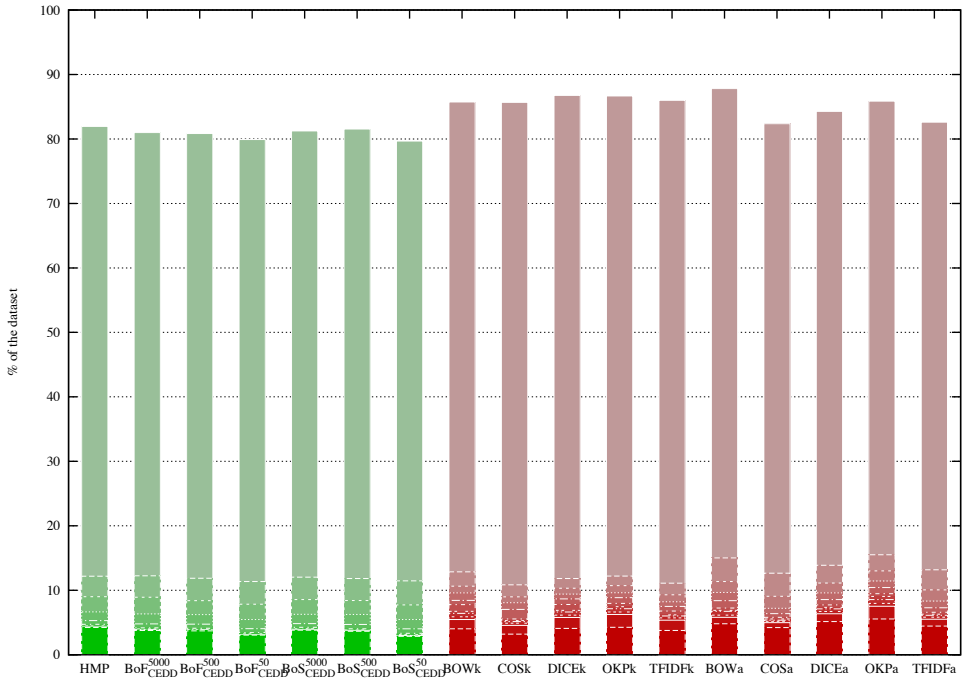
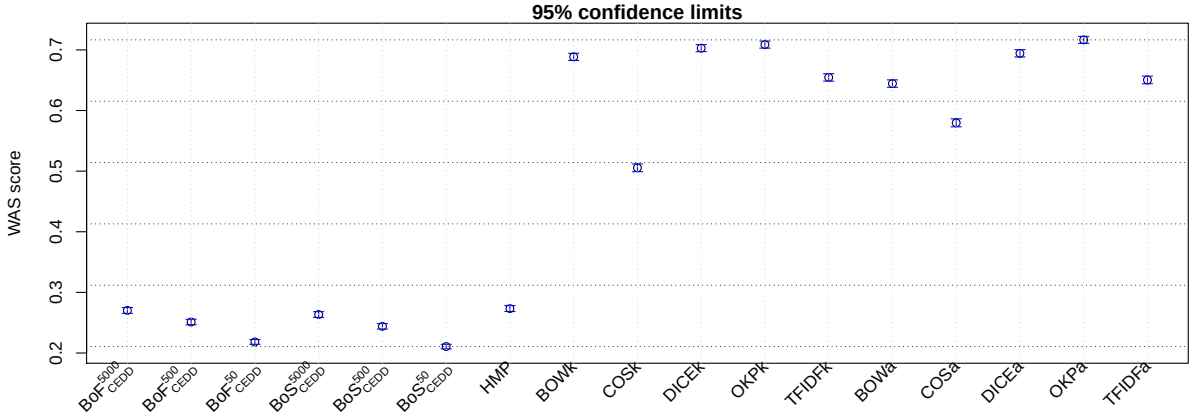


Figure 4.4: Stacked bars showing the isolated performances of each method in the *test* set.

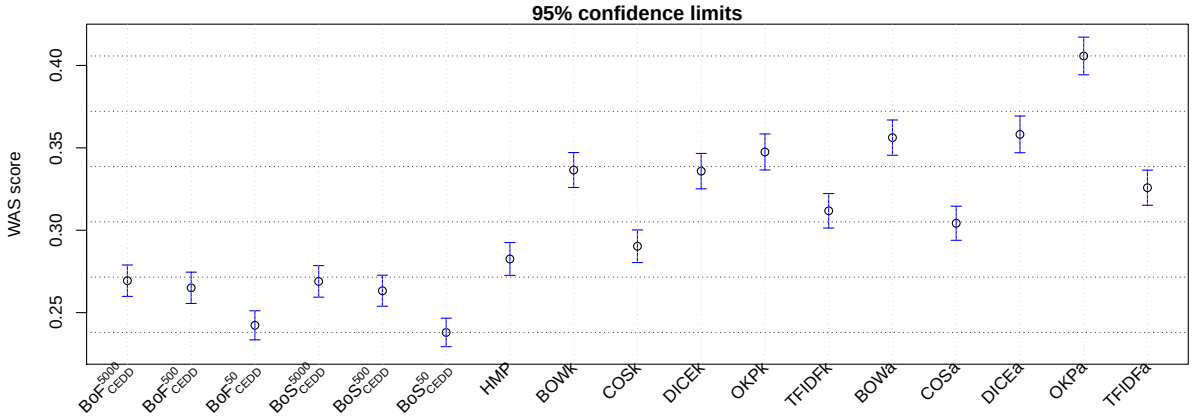
Correlation analysis

This section analyzes the correlation among different features and modalities. Our objective is to assess how those features co-vary with each other. In many situations, the correlation analysis provides additional cues that are very useful to select methods to be combined [108]. We have performed a correlation analysis to evaluate the most promising combinations for the text and visual-based methods. Figure 4.6 shows the correlation graph (from R package *corrgram*) for the development set results. In this case, we consider the distances between a predicted point and its ground truth, that is, we took into account these distances for each query to analyze dependence (Pearson correlation coefficient) of the result using different pair of descriptors.

This kind of plot, aka correlogram, is presented by Friendly [42] and shows the correlation values for each pair of methods as a squared matrix. The darker the color, the higher the correlation value. In the lower triangle, the correlation value is denoted by the intensity of the cell color. In the upper triangle of this matrix, the correlation value is given by the size of the painted area in the circles, as well as by their colors. The diagonal of this matrix holds the name of the methods corresponding to each row and column. Thus to get a sense of the correlation based on painted circles between, for example, BoF^{5000}_{CEDD} versus all the others, we should look at the intersection of cells in the column where it sits



(a) *development* set



(b) *test* set

Figure 4.5: Error bars of WAS(m) measure for isolated methods.

(e.g., last column) with the corresponding row for the others. Conversely, for OKPa (first row) we should look at the intersection of the first row with the corresponding column for each other descriptors.

The correlogram indicates higher correlation among the different textual descriptors, because of the darker color in their cells and the bigger size of the painted area in corresponding circles. The same behavior can be observed for the correlation scores among the different visual descriptors. However, between the textual and visual descriptors, the correlation is very low (the lightest colors and smallest painted areas in their corresponding circle). As we stated before, the best combinations occur when the inputs are independent and non-correlated [27]. Therefore, textual and visual-based methods are very suitable for the combination.

As seen previously, the best geocoding result was yielded by OKPa, so we will analyze the correlation value and the mean WAS of OKPa with each of the other descriptors as

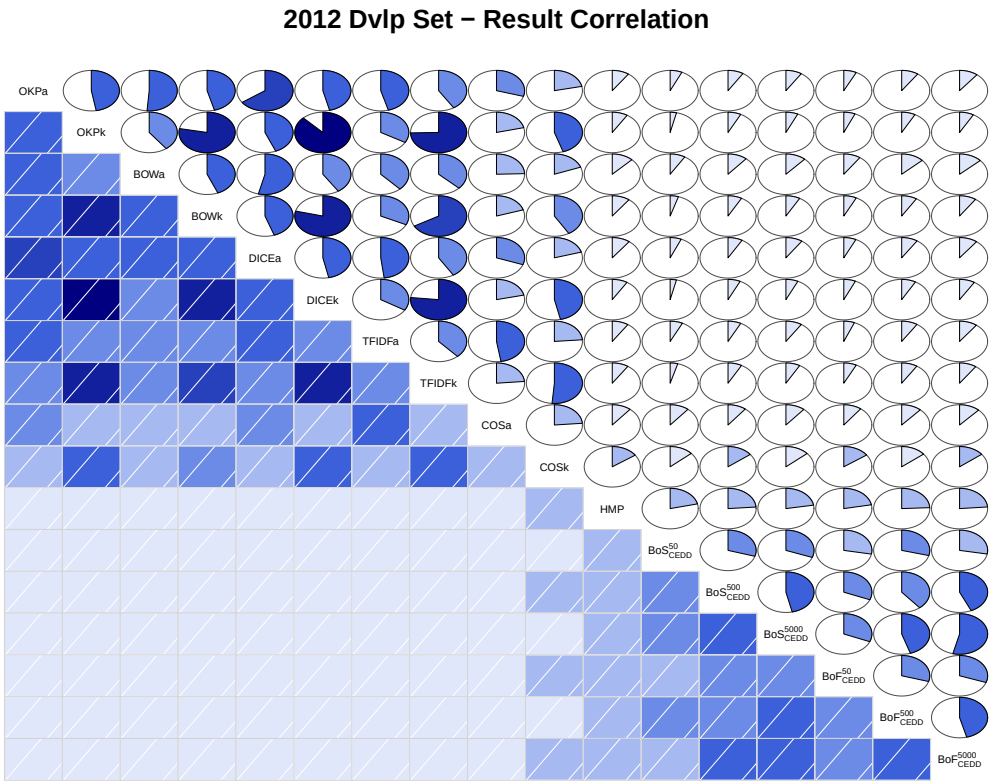


Figure 4.6: Correlation values for each pair of methods evaluated in the *development set*.

depicted in Figure 4.7. We can observe that the best mean WAS score (axis x) is for OKPa and DICEa, followed by OKPa and OKPk. However, the correlation of the first pair is also the highest one (0.6516) while for the second pair it sits in the middle (0.4723), considering the textual pairs correlations (they range between 0.2162 and 0.6516, while their means WAS range from 0.6125 to 0.7140). The combination of OKPa with other visual descriptors results led to the lowest mean WAS concentrated between 0.4650 and 0.4963 (with HMP) and the lowest correlation ranging from 0.0687 to 0.1063.

In the coming section we discuss the results of multimodal combinations and the reasoning behind the combination choices we made based on individual results of each descriptor and their correlation.

Fusion results

The choice of the best textual and visual methods was made based on the correlation analysis of the results of each descriptor. Promising results have been reported using similar approaches [103].

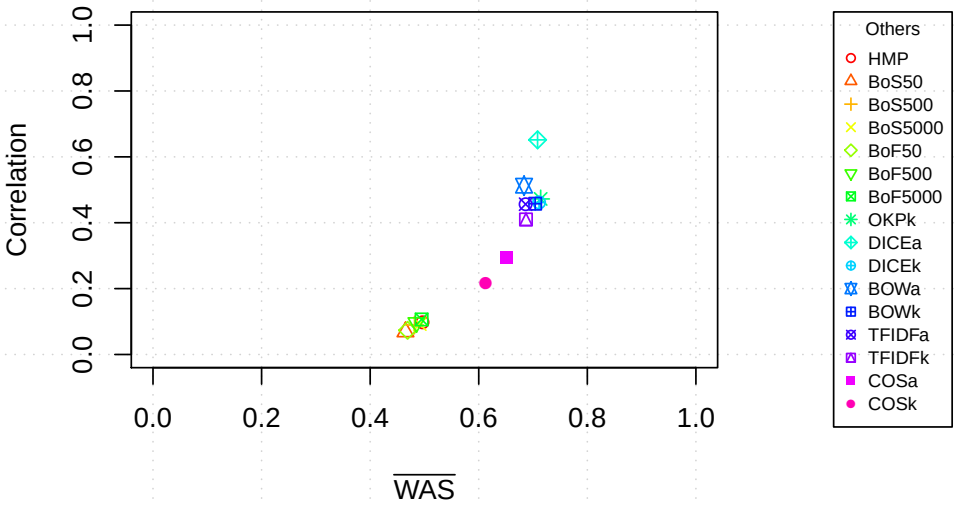


Figure 4.7: Correlation (distance) × average WAS for each pair OKPa vs. other methods evaluated in the *development set*.

Since our work here is focused on data fusion, we will detail our submissions [78] to Placing Task at MediaEval 2012 considering combined results for:

- **only textual** (Ftext) combines results from textual descriptors Okapi and Dice, considering three implementations which yielded the best results (Figure 4.5a) and that have low correlation: Okapi applied to three textual metadata fields (title, description, keywords) associated with a video (OKPa); Okapi applied to the keywords field (OKPk), as well as Dice applied to the keywords (DICEk). These text-based methods have the best scores and are not so highly correlated.

- **only visual** (Fvisual) combines results from the three best visual descriptors as shown in Figure 4.5a: HMP, BoS^{5000}_{CEDD}, and BoF^{5000}_{CEDD}.

- **text & visual** (FTxVis) combines two textual and two visual features: OKPa, OKPk, HMP, and BoS^{5000}_{CEDD}. For textual descriptors, the highest score are the two versions of Okapi and Dice. Looking at the correlation of OKPa (best version) with the other best text descriptors versions, DICEk and OKPk were tied with the lowest correlation (Figure 4.6). Thus, OKPk was paired with OKPa due to its higher score. Using the same reasoning for the visual descriptors, the best ones are HMP, BoF^{5000}_{CEDD}, and BoS^{5000}_{CEDD} with a similar correlation between HMP and the last two. We chose BoS^{5000}_{CEDD} because it is based on Flickr photos, which might be a better match to complement HMP approach (based on videos).

In the next subsection, we evaluate the three rank aggregation methods considering features defined by Ftext, Fvisual, and FTxVis.

Evaluation of Rank Aggregation Methods

In order to choose among the three implemented fusion methods, we used our proposed scoring system WAS to analyze the overall performance. Each of the fusion methods was applied to combined text (Ftext), visual (Fvisual), and text and visual descriptors (FTxVis).

In Figure 4.8, the WAS for each fusion method and their respective error intervals are shown. In the graphic, the result for each fusion method is suffixed with ".M", ".B", or ".R" to indicate that it was generated, respectively, by Multiplication, Borda, or Reciprocal Rank Fusion (RRF) methods detailed in Section 4.1.1.

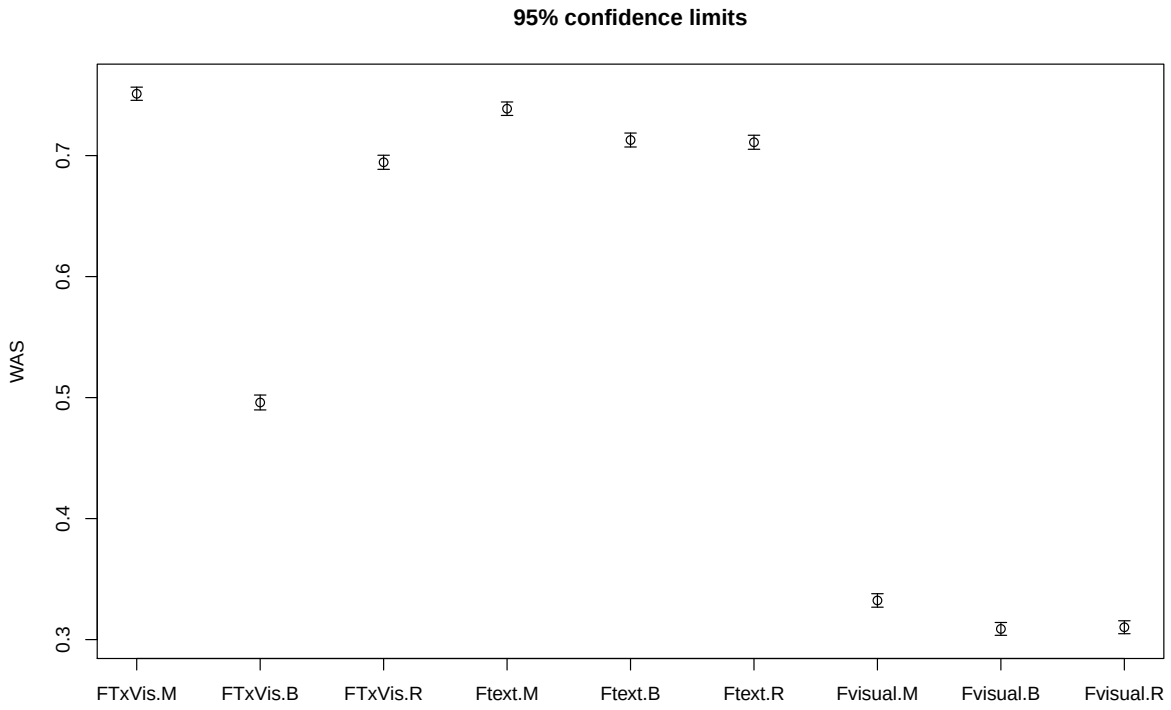


Figure 4.8: Results of rank aggregation methods evaluated using WAS(m) and their standard error (SE) interval.

As we can observe, the multiplication method (FTxVis.M, Ftext.M, and Fvisual.M) yields statistically significant better result with 95% confidence when compared to the other fusion method (no intersection in their confidence interval). Due to these results, from now on we consider the use of the multiplication approach, when we refer to the rank aggregation step of our geocoding framework.

Combined versus Single Modality

Figures 4.9 and 4.10 show the stacked bars comparing the results yield by various methods, for each widening circle used in the Placing Task evaluation, in development and test sets. Those figures show the best methods of each modality (red bars for textual and green ones for visual) used in the combination experiments, as well as the results of their combination (blue bars). Both figures show that fusion methods yield better results than the use of single descriptors (either visual or textual).

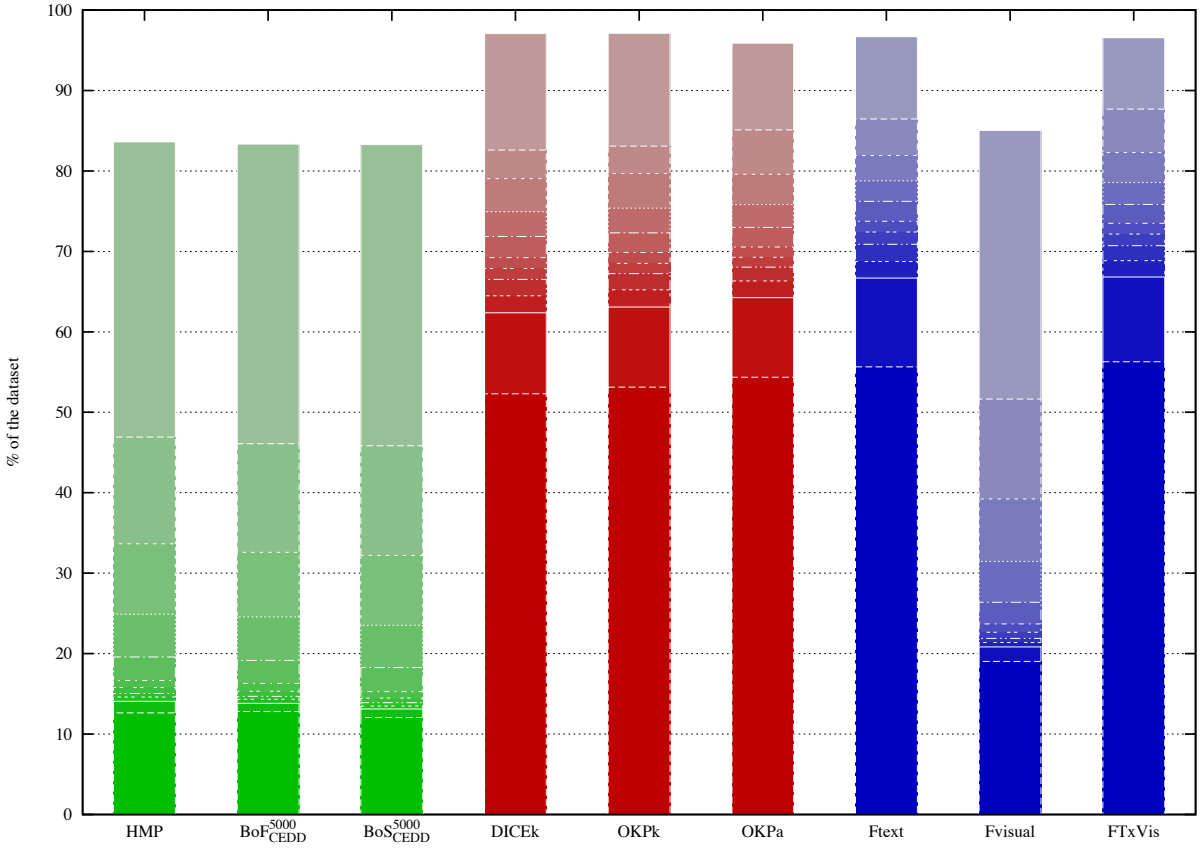


Figure 4.9: Stacked histograms showing the performances, in the *development* set, of the best methods for each modality and their fusion.

Figures 4.11a and 4.11b compare the results of the fusion method that combines textual descriptors with the results of a single feature using our proposed *WAS(m)* score, considering both the development and the test sets. As it can be observed, the fusion of textual descriptors (Ftext) is better (higher score) than the use of features in isolation (OKPa, OKPk, and DICEk), in both development and test sets, with a statistical significance of 95% confidence limit.

Figure 4.12a shows the results for the combination of visual features (Fvisual). It

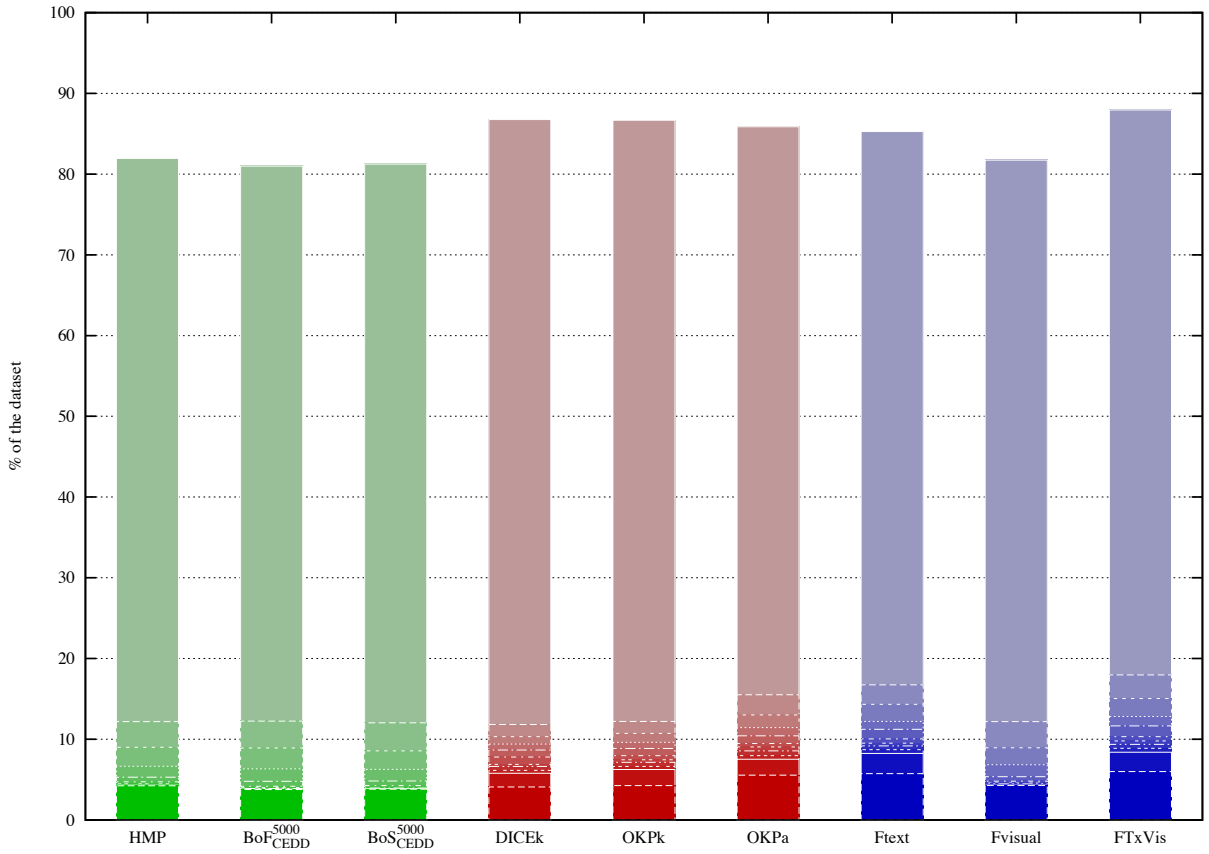
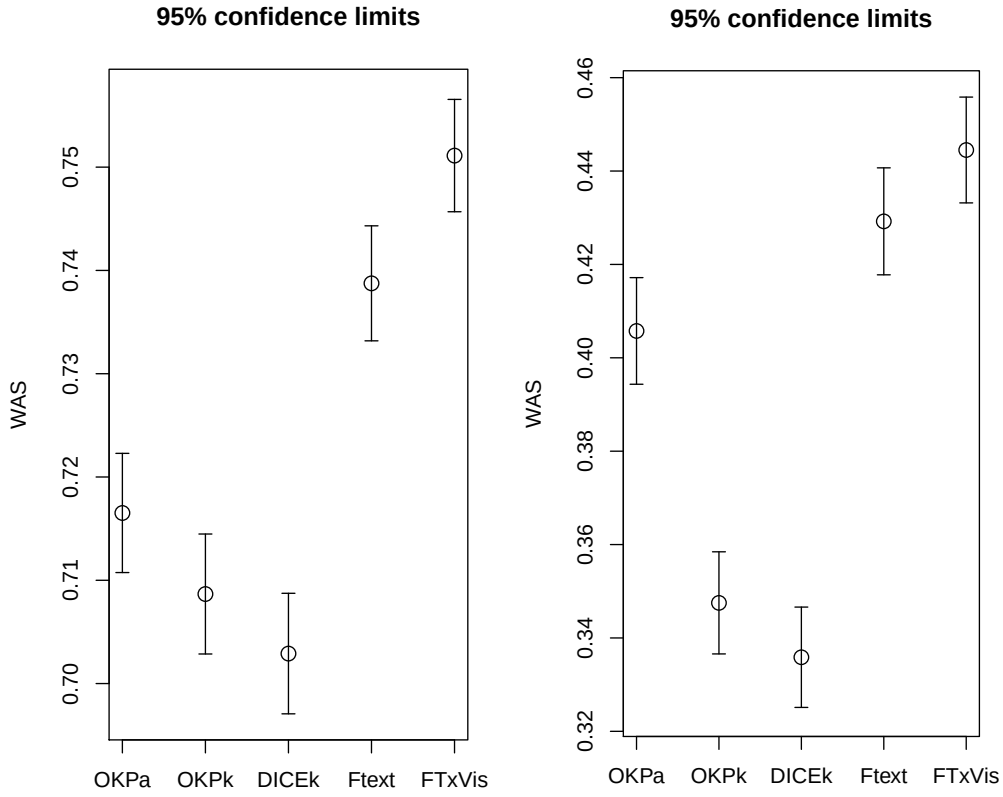


Figure 4.10: Stacked histograms showing the performances, in the *test* set, of the best methods for each modality and their fusion in the test set.

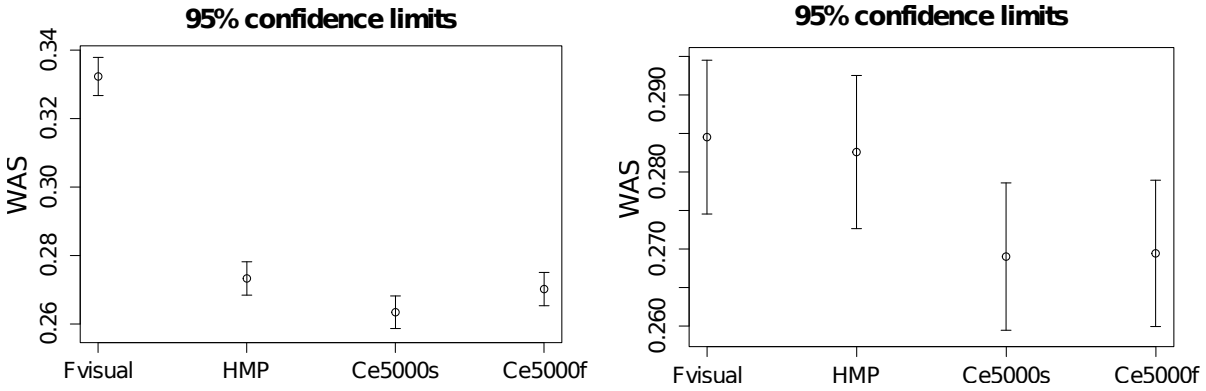
shows that the combination of HMP, BoF^{5000}_{CEDD}(Ce5000f), and BoS^{5000}_{CEDD} (Ce5000s) in the development set improved significantly (95% confidence limits) over the best visual individual result (HMP) in the development set (0.3323 against 0.2733). However for the test set, no statistical difference is identified, as shown in Figure 4.12b (0.2845 over 0.2826).

The fusion taking account visual and textual features (FTxVis) also yields better results than the use of a single modality (OKPa) as shown in Figure 4.11. Additionally, the improvement of FTxVis over Ftext is more visible for the development set (0.7511 over 0.7388) than for the test set (0.4445 over 0.4292). However, in the training data set, Fvisual and FTxVis results present statistically significant difference, as shown in Figure 4.11a, while Figure 4.11b shows that in the test set their results do not. One of the reasons might be due to the use of the same weight for combining textual and visual features. In addition to this, there were more videos without textual data to help the geocoding in test set than in training set.



(a) development set (b) test set

Figure 4.11: WAS(m) general score and standard error (SE) interval: fusion and individual *textual* descriptors results in the *development* set (a) and *test* set (b).



(a) development set (b) test set

Figure 4.12: WAS(m) general score and standard error interval: fusion and individual *visual* descriptors results in the *development* set (a) and *test* set (b).

For the concatenation of title, description, and keywords, in development set, there was no textual data in approximately 1% of its videos, whereas its was the case for about 7% of the videos in the test set. When considering the keywords alone, these percentages increase: around 11% and 35% respectively.

Consider the case in which the textual approach provides a perfect estimation and the visual method performs an incorrect prediction. Once the textual and visual features have the same contribution in the final result, their combination may not improve the overall performance of individual strategies. Therefore, there is room for improvements in the fusion module, by incorporating new strategies for assigning different relevance weights to each method being combined, for example. We will address this research venue in future work.

In summary, we can see better results when combining methods of different modalities or descriptors. The fusion of the three text methods (Ftext), as well as the fusion of visual and textual descriptors (FTxVis) overcome the best single descriptor method (OKPa), as shown in Figures 4.11a and 4.11b.

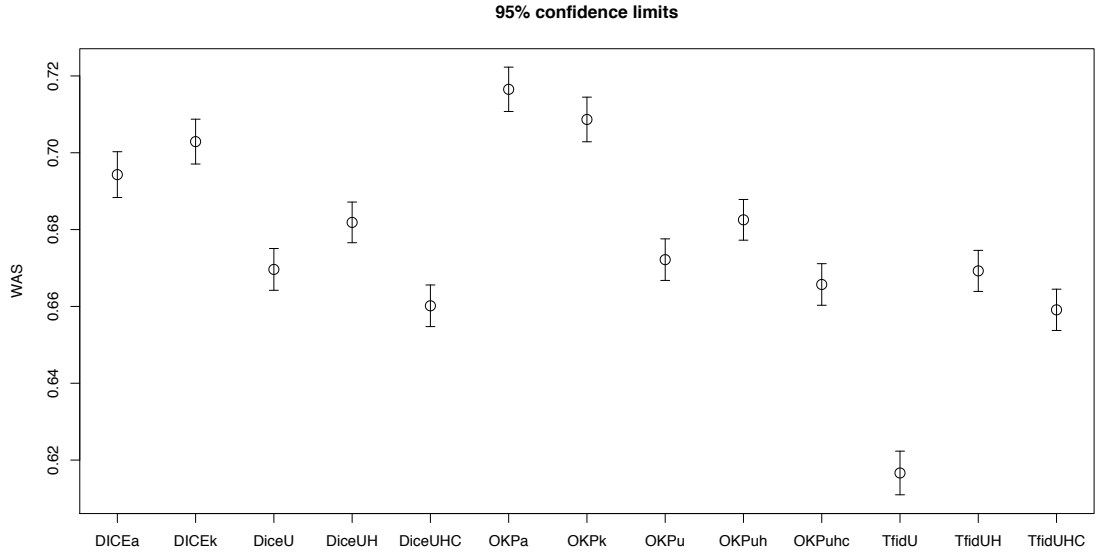
Incorporating User-Related Data

This section describes experiments to evaluate the impact of using user-related data. We analyze different geocoding strategies based on combining our best selected visual and textual features with ranked lists defined in terms of (U) just user names found in the videos' metadata, (UH) user names and the videos' owner declared home location, and (UHC) the concatenation of user names, home location, and comments related to each video.

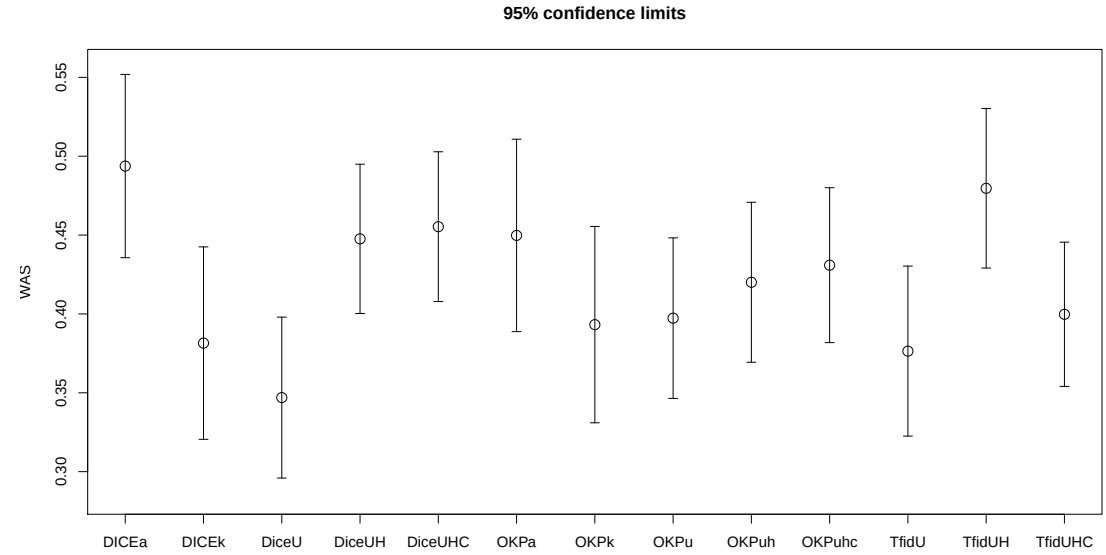
We treat the user-related data as another textual information. We used the textual descriptors we described in Section 4.1.1 to index and process them. We also compare our best single "conventional" textual results (DICEa, Dicek, OKPa, and OKPk) with the different strategies to incorporate user information (U, UH, and UHC).

Figure 4.13 shows the results, for the training and test sets, in term of WAS scores and confidence intervals. Notice that, in the training set (a), features based on user information yields worse results when compared to the "conventional" textual features. The best results for user-related features are observed for DiceUH (0.6819), OKPuh (0.6825), and TfIdUH (0.6692), that is, the geocoding strategies that consider user names and owner location (UH). For the test set (b), all the user-related features results are statistically similar (there are intersections among their confidence limits), although worser scores (WAS) were observed for the test set when compared with those for training set (a). This is also true when comparing them to "conventional" textual features.

We also conducted correlation analysis in order to decide which methods for user-related data should be used in the rank aggregation module, starting from their top three



(a) development set



(b) test set

Figure 4.13: Geocoding results for ranked lists defined by conventional textual (using OKPa, OKPk, DICEa, and DICEk similarity functions) and user-related properties (using OKP, Dice, and TfIdf).

results. Figure 4.14 shows that user-related features are low (light color) correlated with both conventional text and visual descriptors (HMP and Ce5000s), which indicates that better results can be produced when combined. Regarding user-related feature, we used as starting point the descriptor with higher WAS, which is OKPuh. Thus, we choose to pair it with TfIdUH as OKPuh is lower correlated with this than with DiceUH.

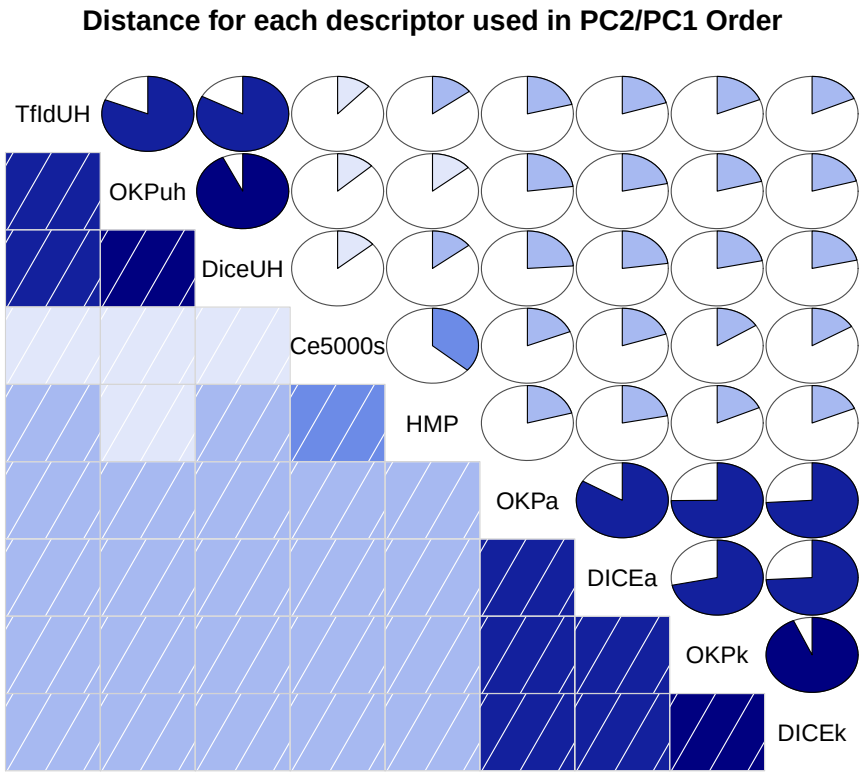


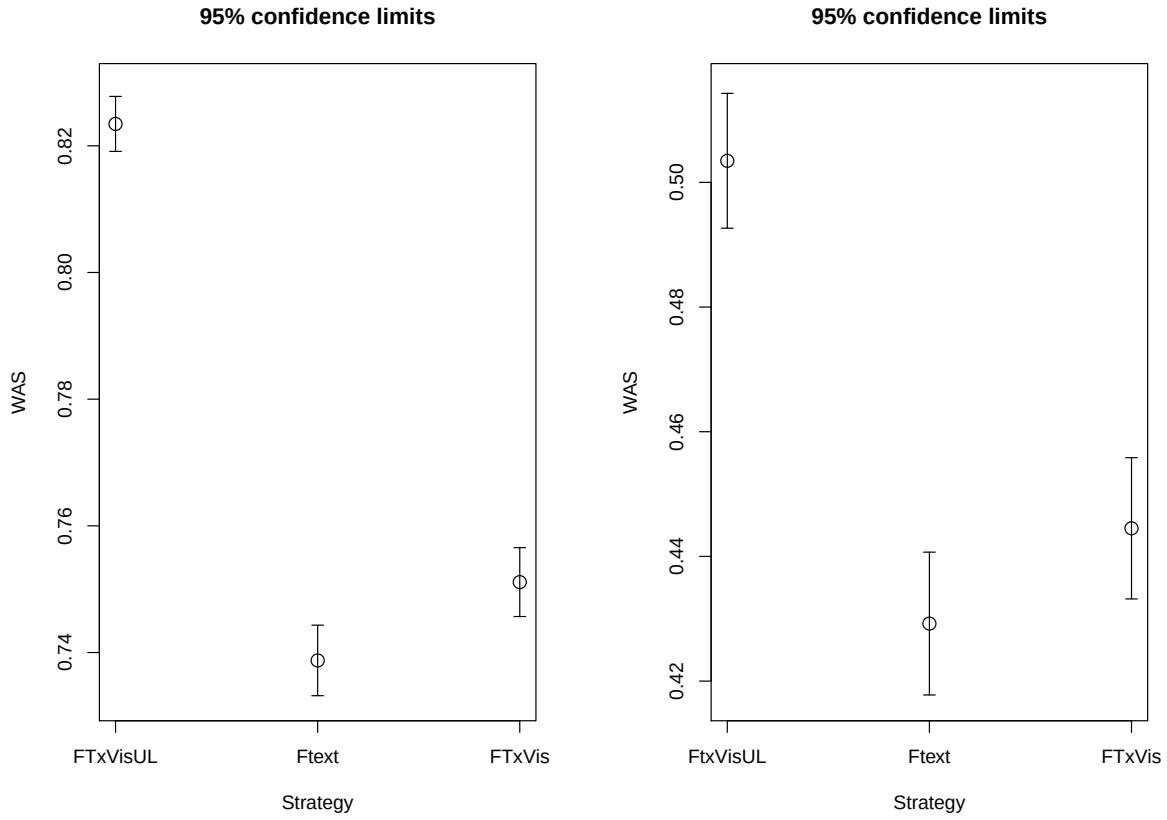
Figure 4.14: Correlogram in the development set for conventional text (OKPa, OKPk, DICEa, and DICEk), user-related (TfIdUH, OKPuh, and DiceUH) features, and two best visual features (Ce5000s and HMP).

Figure 4.15 shows the geocoding results of three different strategies: Ftex, which considers the three best textual features; FTxVis, which uses the best two visual and textual features; and finally TxVisUL, which combines the two best visual, textual and user-related features. These results consider the use of the multiplicative rank aggregation method. As expected, the user-related features improves the geocoding results in both development and test sets. Their results are significantly better than of the other strategies, with 95% of confidence.

These outcomes support our hypothesis that fusing results of different modalities can improve the final geocoding results.

Comparisons with other Video Geocoding Initiatives

This section compares the proposed geocoding method with the ones provided by other participants of Placing Task 2012. We have not used WAS here as the evaluation measure,



(a) development set (b) test set

Figure 4.15: WAS(m) general score and standard error interval: fusion results for three different geocoding strategies (Ftex, FTxVis, and TxVisUL) in the *development* set (a) and *test* set (b).

since distance scores for each query video are not available for the geocoding methods defined by other participants. The organizers of Placing Task compare primarily the results reached within 1 km radius, although results in other precisions are shown as well.

First, we compare the submissions that only consider the use of *visual* content. As it can be observed in Figure 4.16, our results are the best at 1 km precision (15.93%). Our solution refers to the combination of our best visual descriptors results (Fvisual). Note that other teams only achieve this level of accuracy for 1,000 km radius. In fact, even at this radius, our method (UNICAMP) is still ahead, with 25.47% of test videos being correctly geocoded.

Figure 4.17 presents the best-performance (external information allowed) of all participants, at 1 km radius. Our results (UNICAMP) consider geocoding strategy based on combining visual and textual descriptions, but **without** using any user-related data.

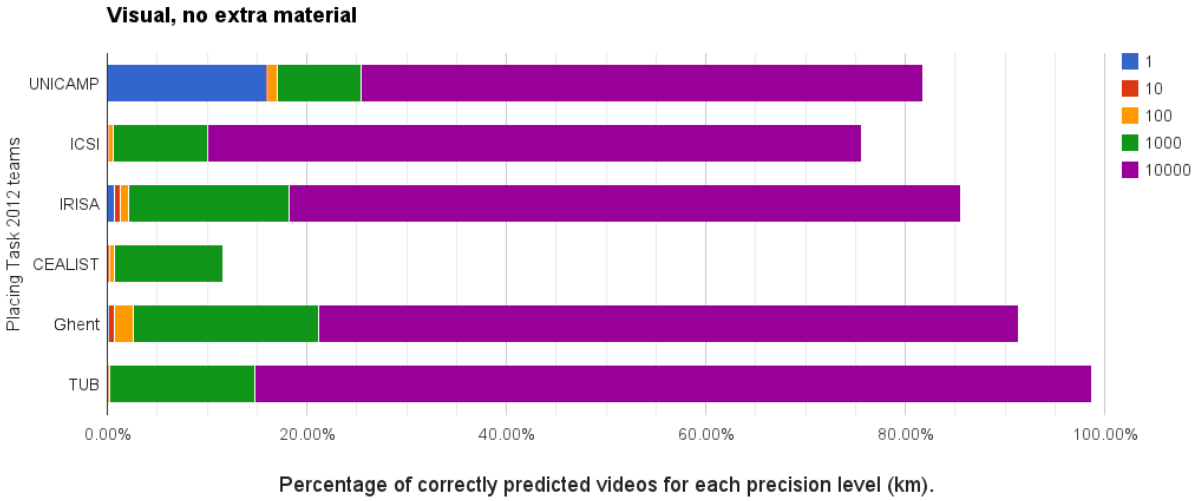


Figure 4.16: Only-visual submission: correctly geocoded test videos for different precision levels.

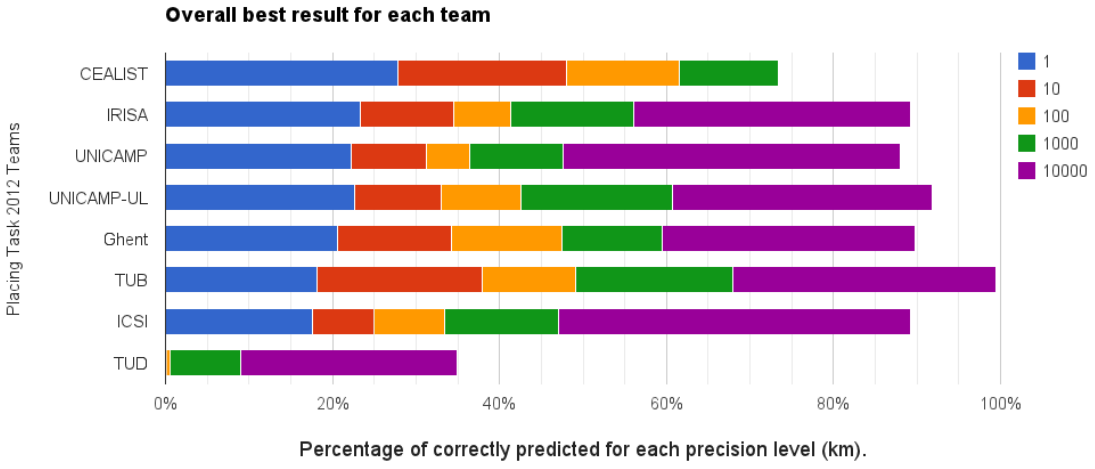


Figure 4.17: Overall best submission to the Placing Task 2012, considering correctly geocoded test videos within different precision radii.

Although UNICAMP achieved the third place in the overall best performance, results show how promising our proposed multimodal framework is to geocode videos. We have even outperformed some other submissions that used extra information. Our textual and the visual processing modules use straightforward information retrieval techniques (KNN searches) to query the test videos against the development set, and then to assign location. The other participants of Placing Task at MediaEval 2012, on the other hand, implemented *ad hoc* methods to define the location for a video (Section 2.2). The CEALIST team (first place), for example, exploits additional information, not provided by organizers, to learn tagging patterns of users to improve its results.

Figure 4.18 shows the same results, but now eliminating those submissions that used additional information. In this figure, we also consider the implementation of our framework with two different strategies based on the incorporation of user-related data. UNICAMP results are the same reported in Figure 4.17 and does not include user-related data. UNICAMP-UL, in turn, stands for the strategy that considers user-related data. As it can be observed, UNICAMP-UL yields comparable results to the best submissions (IRISA1) that do not use additional/external resources.

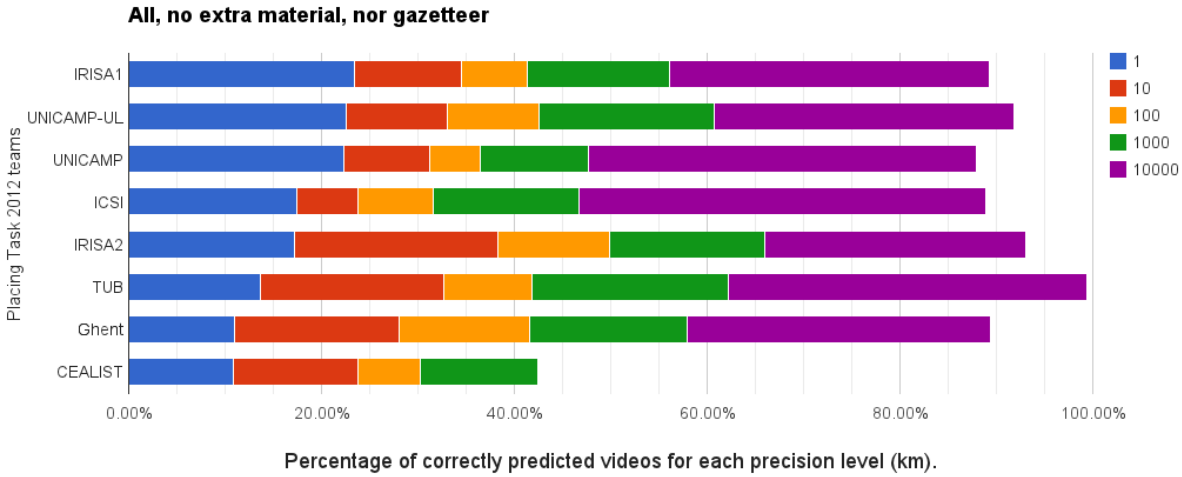


Figure 4.18: Effectiveness performance for different precision levels. Results for methods that do not use any additional resources, nor gazetteers (ordered by results for 1 km precision).

4.2 Domain-specific Image Geocoding: Virginia Tech Building Photos Case

As presented before, the Crisis, Tragedy, and Recovery network (CTRnet) project collects news and online resources (web pages, public Twitter and Facebook posts) related to natural disasters and man-made tragedies.

To support the creation of map-based browsing services building on photo content, the first step is to be able to geocode images. For this, we need to evaluate image descriptors in image geocoding tasks.

In [105], the authors explored a strategy based on global descriptors to tackle the challenge of geocoding videos based only on visual features. In more recent studies [78, 84], we combined visual and textual features for video geocoding. We would like to explore similar approach in the context of Virginia Tech Building Photos.

The objective now is to investigate the most suitable image descriptors to be used to geocode photos belonging to the collection related to the VT April 16 shooting event. Later, geocoded photos could be used to create map-based photo browsing services.

The first insight to tackle the problem of geocoding photos about VT is to leverage research to recognize buildings. Previous initiatives, such as [107], perform matching of local descriptors to find similar regions within images of a set of buildings. Although they are not explicitly geocoding images, their approaches could be used for that purpose. They worked with buildings from the University of Oxford.[7] After describing images with a scheme based on a visual vocabulary (quantized local features), matching was done between a given query image and images from the dataset. Performance was compared for different vocabulary sizes, as well as vocabularies generated by diverse methods. Here we will employ a similar strategy, with the aim of evaluating the performance of local image description approaches in the task of geocoding building photos.

For this geocoding experiment, we need a data collection and the evaluation criteria which will be described below.

4.2.1 Datasets

We use two datasets in our experiments. One is used as our visual knowledge base and will be referred to as training data. The other includes the test data images whose locations will be predicted by the proposed geocoding system.

Training dataset The training data is a subset of 4,852 photos from VT's University Relations (UniRel) Photo Library. Each photo has some metadata associated, such as

[7] `http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/` *(as of Dec. 2013).*

keywords, caption describing the scene, date, camera model, and photographer's name. For our purpose, we filtered the photos by the content of keywords and caption fields. As we were interested in the university buildings, we searched for photos whose metadata (keywords or caption) contains building and place names (e.g., Duck Pond). The building/place names list was built up from both the VT site[8] and the campus building database maintained by GIS staff for campus facilities. The resulting training set contains photos of buildings or places with their location. Figure 4.19 shows the spatial distribution of buildings whose photos are in the training set.

Test dataset The test dataset[9] contains 565 photos of VT buildings. Most of them were obtained from personal collections while some others were downloaded from the VT website. The photos were obtained under different angle and light conditions. The locations of these photos are shown in Figure 4.20. Note that the test set covers a smaller area (near to the Drillfield in the campus center) when compared to the training set.

Figure 4.19 and Figure 4.20 were generated using a tool provided by sethoscope.net.[10] We used the option that employs tiles, provided by Stamen Design, under CC BY 3.0. Data was provided by OpenStreetMap, under CC BY SA.

Ground truth The ground truth for the images in the training and test data sets, that is, the "correct" location for each of them, was inferred from the corresponding building/place name associated with the photo. For the training photos, we used the place/building name that appears in their metadata. For the test photos, we use the name that we manually labeled each photo. The ground truth for these photos is based on the latitude and longitude from the VT site, as well as on the result of processing building names by Google's geocoding service. However, if no matches were found by the geocoding service or if disambiguation was needed, the place/building was manually located and confirmed in Google Maps or Open Street Map using its name. Additionally, some photos and some of the resulting geocoding locations were visually and manually inspected to determine their final location and/or coordinates.

The tool we used in this process is geopy,[11] a geocoding toolbox for Python that accesses popular Web geocoding API services. Geopy also supports the computation of geographic distance between two given lat/long points. We used this feature to evaluate our geocoding results.

[8] `http://www.vt.edu/about/buildings/` *(as of Dec. 2013).*
[9] `http://www.recod.ic.unicamp.br/VTBuildings/` *(as of Dec. 2013).*
[10] `http://www.sethoscope.net/heatmap/` *(as of Dec. 2013).*
[11] `https://github.com/geopy/geopy#readme` *(as of Dec. 2013).*

Figure 4.19: Spatial distribution of photos used as *training set*.

Figure 4.20: Spatial distribution of photos used as *test set*.

4.2.2 Evaluation Criteria

The evaluation criterion used here is inspired by the evaluation procedure adopted in the Placing Task at MediaEval [112]. The effectiveness of a method is based on the geographic distance (great-circle distance) of the estimated geo-coordinates of a digital object to its corresponding ground truth location, in a series of widening circles of radius. An estimated point is counted as correct if it is within a particular circle size, that is, a radius value or precision level.

In our case, we are interested in determining as accurately as possible the location for a photo image. Furthermore, our area of interest is restricted to the Virginia Tech campus, so our precision level should be in the range of meters. Taking into account that the two farthest points of the town of Blacksburg (where VT is located) are about 10 km apart, we can accept that two points on the VT campus should not be further apart than 5 km. The precision levels adopted are {1, 50, 100, ..., 1000}, {1100, 1200, ..., 2000},

and {3000, 4000, 5000} meters.

4.2.3 Setup

First, the visual content properties of each provided image are encoded into feature vectors, considering all evaluated descriptors. Then, the visual distances between the photos in the test set and all photos in the training set are computed. Finally, for each test photo, a ranked list of training photos is produced.

To represent each image, we used the bag-of-visual-words model [120]. In that model, after extracting low-level features with local descriptors, we quantize the feature space in order to obtain a visual dictionary (codebook) and then we represent each local description according to the dictionary. For low-level feature extraction, we used: dense SIFT (6 pixels) [124], sparse SIFT (Harris-Laplace detector) [124], and sparse SURF (Fast-Hessian detector) [6]. We randomly quantized the feature space [127], generating two dictionary sizes: 1,000 and 10,000 visual words.

To compute the bag-of-word representation, we used soft assignment (σ=60 for SIFT and σ=0.08 for SURF) [125] and two pooling methods: max pooling [12] and Word Spatial Arrangement (WSA) [106]. WSA was used only over the sparse SIFT, while max pooling was used for all low-level features. Table 4.1 lists the evaluated methods.

Table 4.1: Image representations evaluated.

Acronym	**Method**
D.SIFT.1k	dense SIFT, 1,000 words, soft assignment (σ=60), max pooling
D.SIFT.10k	dense SIFT, 10,000 words, soft assignment (σ=60), max pooling
S.SIFT.1k	sparse SIFT, 1,000 words, soft assignment (σ=60), max pooling
S.SIFT.10k	sparse SIFT, 10,000 words, soft assignment (σ=60), max pooling
S.SURF.1k	sparse SURF, 1,000 words, soft assignment (σ=0.08), max pooling
S.SURF.10k	sparse SURF, 10,000 words, soft assignment (σ=0.08), max pooling
WSA.1k	sparse SIFT, 1,000 words, soft assignment (σ=60), WSA
WSA.10k	sparse SIFT, 10,000 words, soft assignment (σ=60), WSA

Geocoding Process The geocoding scheme adopted is based on performing K-nearest neighbor (KNN) searches. In this study, the location of a test photo is defined based on the geographic coordinates of the most similar image in the training set, i.e., is defined in terms of the location of the 1-nearest neighbor (i.e., $K = 1$) of the test photo. The visual distances between an input test image and all training images are computed. Training images are then ranked in ascending order of their visual distance to the input test image,

and the latitude and longitude coordinates of the top-ranked training photo are assigned to the test image.

4.2.4 Results

Test Results for single feature

Figure 4.21 presents the geocoding results for evaluated methods, considering different precision levels. Observe that the S.SURF.10k descriptor yields better results starting from the 150 m precision level on, followed by the D.SIFT.1k descriptor. Given that there are neighboring buildings in VT that are apart from each other (measured from their centroid) about 100 to 200 meters, it is reasonable to tolerate a maximum estimation error around 200 m. The S.SURF.10k method geocoded correctly 20% of the photos within 150 m of the actual location and almost 80% within 600 m. For the precision level of 1 m, WSA.10k geocoded more images.

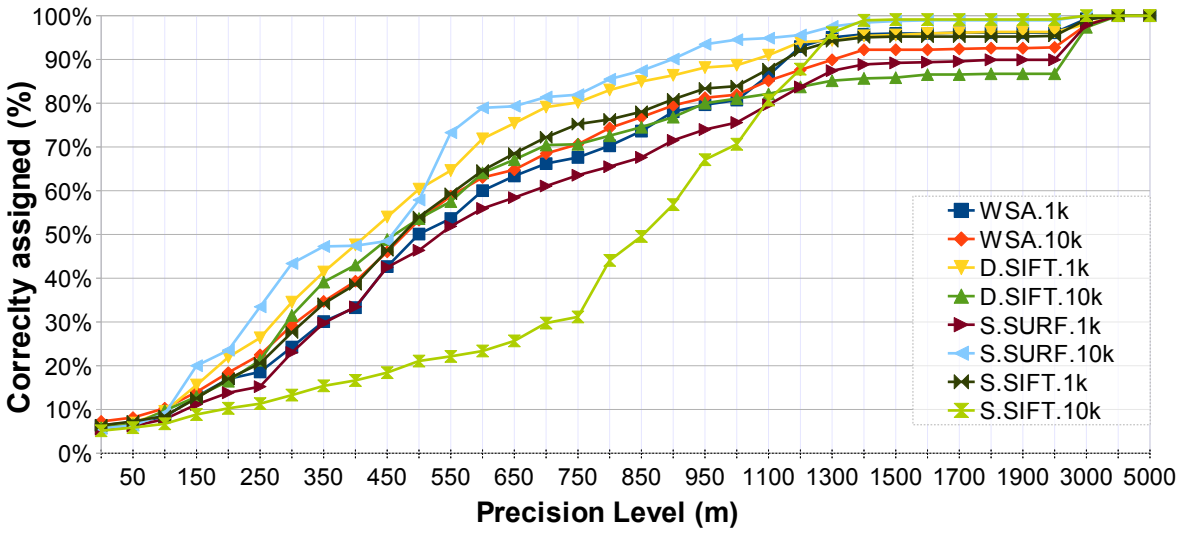


Figure 4.21: Correctly predicted test photos.

Examples of geocoding results

Table 4.2 shows some examples of query photos of the test set and the corresponding top-similar images in the training set for each visual descriptor. Each table cell also presents the geographic distance between the top-ranked training image and the query photo's ground truth.

Table 4.2: The best visual match for each query image and its geocoding result. Values below the photo thumbnail show the geographic distance (in meters) to the location of query image.

Query	Building Name	D.SIFT.1k	D.SIFT.10k	S.SIFT.1k	S.SIFT.10k	S.SURF.1k	S.SURF.10k	WSA.1k	WSA.10k
	Holtzman Alumni Center								
M_holtzman		828.68	677.83	521.63	1444.40	3097.05	1213.12	3097.05	3097.05
	Major Williams Hall								
M_major-williams		344.41	344.41	261.14	878.87	344.41	293.99	874.65	874.65
	Lane Hall								
P1080012		424.08	86.87	349.62	161.13	217.67	1231.06	238.44	238.44
	Shultz Hall								
P1080060		577.29	73.81	341.67	927.37	273.69	1416.96	1134.17	255.78
	Major Williams Hall								
P1080175		102.50	102.50	874.65	878.87	1246.30	209.30	344.41	245.95
	Davidson Hall								
P1080241		494.87	301.03	273.88	1019.67	252.30	2721.11	578.88	460.76
	Squires Student Center								
P1080509		314.82	516.54	2788.72	1077.80	2788.72	130.10	0.00	526.03
	Torgersen Bridge								
P1080710		906.41	2801.90	74.33	1070.43	0.00	0.00	275.15	74.33
	Newman Library								
P1080711		407.52	370.97	299.52	299.52	299.52	80.13	547.67	869.76
	Pamplin Hall								
P1080371		127.52	212.48	483.67	127.52	127.52	501.05	253.25	403.00

Consider, for example, the top-ranked image in the case of query P1080710 (picture of the Torgersen Bridge). The S.SURF (1k and 10k) descriptor is able to match it to a photo that only pictures a detail of that building, whereas WSA.10k and S.SIFT.1k match that to a photo from the same building but under a different light (darker) condition. However, as this photo was labeled as Torgersen Hall instead of Torgersen Bridge (part of Torgersen Hall), its geographic distance was not zero. The query P1080012 (Lane Hall) shows an example where S.SURF.10k performed very badly. D.SIFT.10k, on the other hand, matched it to a photo of a building (Shanks Hall) that is close (86.87 m) to Lane Hall, while S.SIFT.10k found the query similar to a picture of Torgersen Bridge (161.13 m away) at night.

4.2.5 Feature Fusion

Training set Results and Discussions

In Section 4.1, we have shown that combining individual non-correlated descriptors may improve geocoding results. A correlation analysis helped to evaluate the most promising descriptors to be combined. In order to do that in this context, we will analyze the results for each descriptor evaluated on the training set. For the training set, we will perform experiments considering each image of the training set as a query photo. In this case, given that the query photo always is the best match to itself (thus it will be the first in this list), we use the second photo of available ranked lists to define the final location. As it is stated in Section 4.1.3, this can be seen as leave-one-out cross-validation [46, sec. 7.10.1, p. 242], in which each time a different item in the training set is left out and used as a query against the others in that set.

Figure 4.22 shows the correlation graph (*corrgrams* R package) for the results of the training set. In this case, for each method and query image, its geocoding result is the geographic distance between a predicted point and its ground truth. Thus, we studied the correlation of these results for the evaluated methods. As we can observe, the lowest correlations are among S.SIFT.1k and the others, which is indicated by the lightest colors in the first column and the smallest painted area in circles in the first line.

Figure 4.23 shows the geocoding results of evaluated descriptors on the training set. We can observe that WSA.10k and S.SIFT.1k yield the best performance, followed by WSA.1k and D.SIFT.10k at 200 km. On the other hand, S.SIFT.1k is less correlated with other descriptors (see Figure 4.22), which suggests that its combination with other descriptors may improve the geocoding results.

Comparing the geocoding results on both test and training sets, we found some surprising results. S.SURF.10k, for example, yields the worst results in the training set, but performs very well on the test set compared to the other methods. One possible explana-

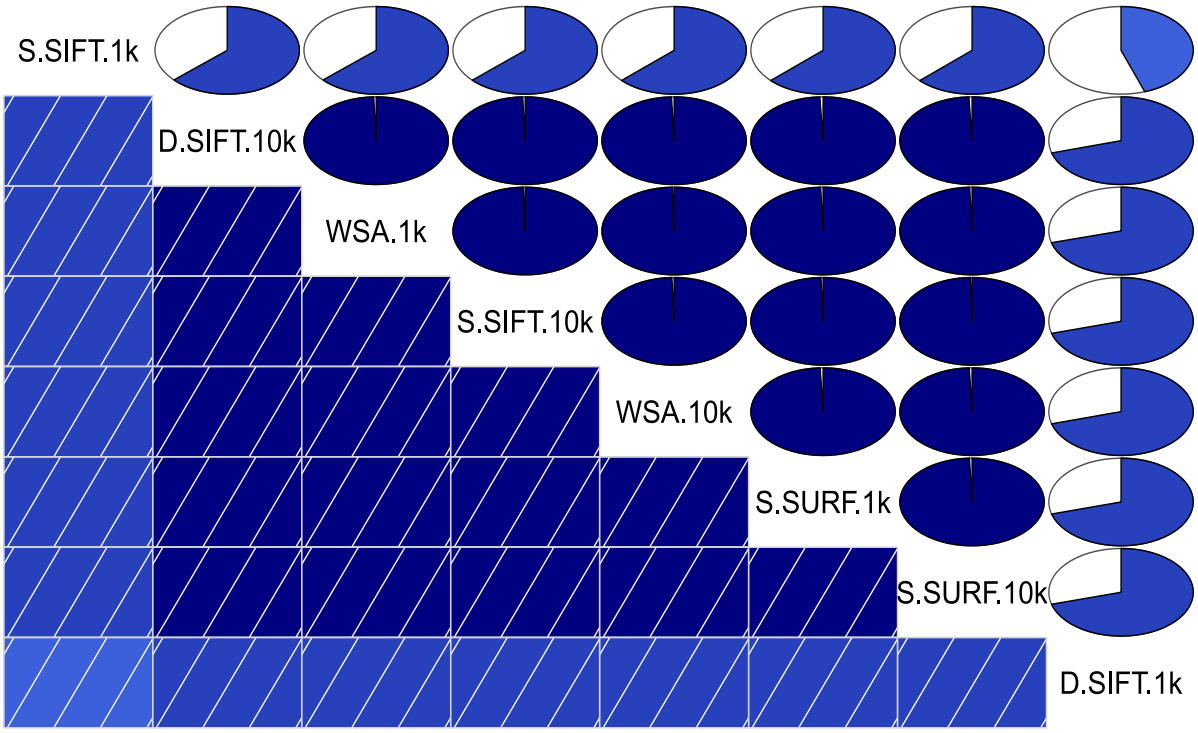


Figure 4.22: Correlation among evaluated descriptors in the *training set*.

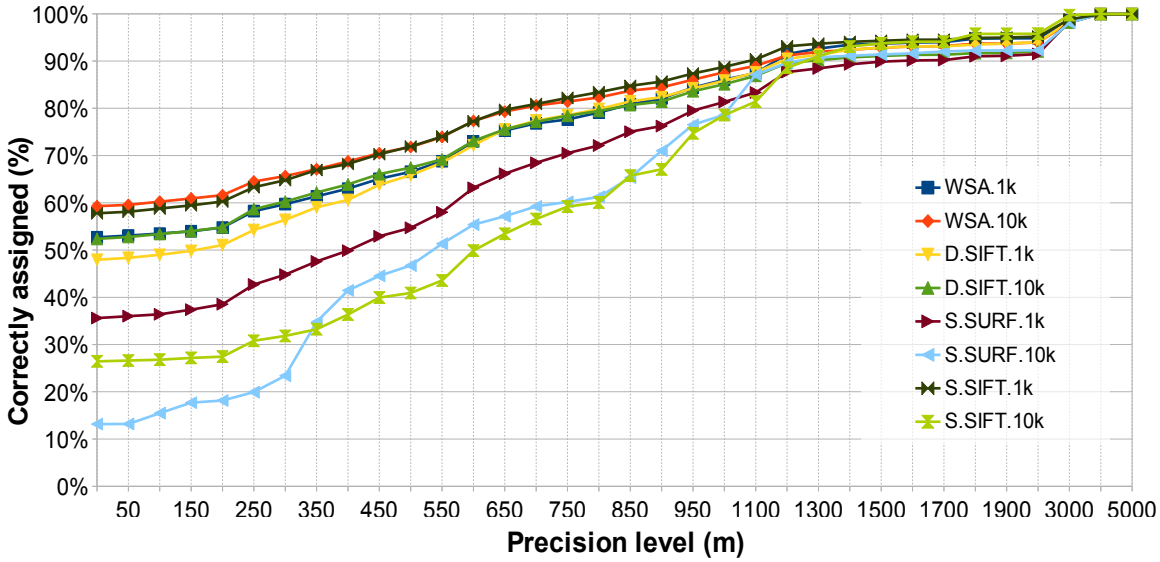


Figure 4.23: Correctly predicted training photos.

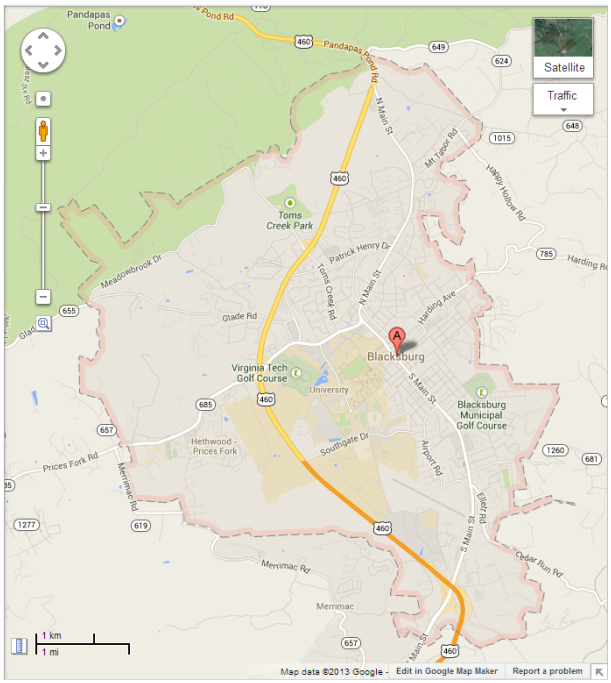


Figure 4.24: Boundary of Blacksburg (VA), where the main campus of VT is located (*Source: Google Maps*).

tion lies on the differences between test set and training set images. In the test set, there are more close up photos, whereas the training set includes pictures in a wider frame, i.e., the images depict more distant objects (buildings).

Until now, we have presented the results using the placing task style evaluation, that is, the accumulative counts of correctly assigned test images for certain precision levels. However, we have proposed another scoring system to evaluate the geocoding results, using a modified version of the WAS measure. We changed a parameter for WAS score computation in order to fit this particular case study. As displayed in Equation 3.2, the factor R_{max} is responsible for making the score ranges in the interval [0,1]. In that section, we have assumed that $R_{max} = 20,027.5\ km$, which is the maximum distance between two farthest points on the Earth. Nonetheless, in this case study, we deal with a smaller region, Blacksburg (VA), the town where VT main campus is located. Therefore, we propose to adjust R_{max} to $11,000\ m$, which is approximately the radius of the minimum rounding circle around Blacksburg (Figure 4.24).

Figure 4.25 shows, for each descriptor, the WAS results with their confidence interval (error bars graphic generated by R, *psych* package). We can observe that the best results in the development set were yielded by S.SIFT.1k and WAS.10k descriptors with no significant difference between them. The second best results are for D.SIFT.10k and

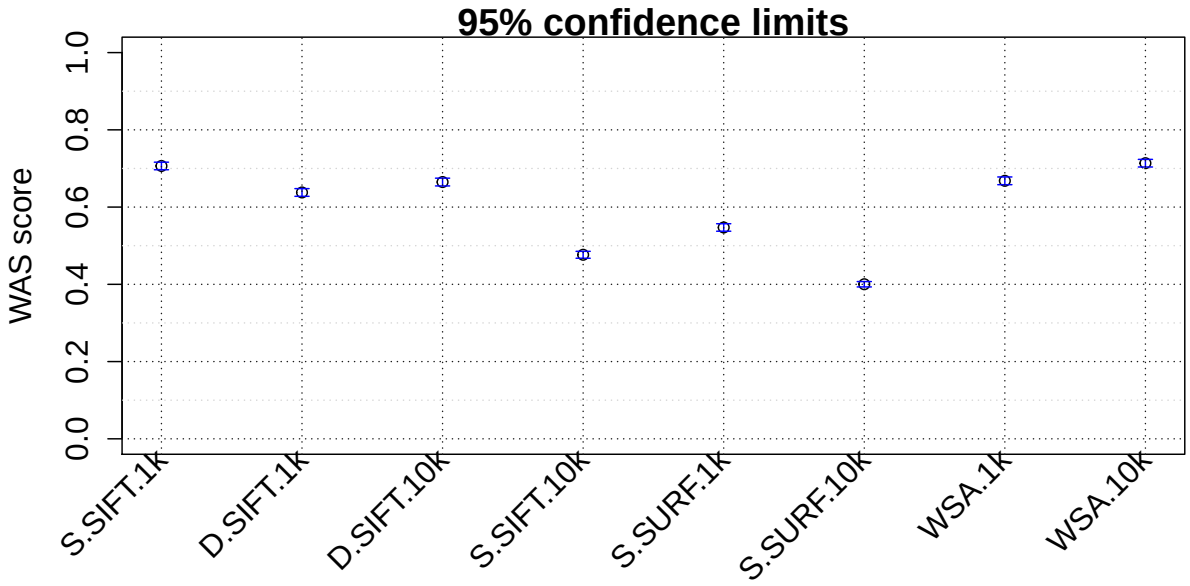


Figure 4.25: WAS scores and confidence intervals for single results in the *training* set.

WSA.1k (also tied), followed by D.SIFT.1k, S.SURF.1k, S.SIFT.10k, and S.SURF.10k. As we can observe, this order of the best results (and ties) is exactly the same for the curve of accumulated count (higher to lower) at 200 m, which is consistent with the results shown in Figure 4.23.

As observed early in Figure 4.22, S.SIFT.1k has the least correlated results with respect to the other's results. It has also yielded the best geocoding results in the training set. Thus, S.SIFT.1k is a good candidate for being combined with other descriptors. Our hypothesis is that the use of the pair of descriptors with the highest mean WAS and the lowest correlation will potentially improve the geocoding result.

Figure 4.26 shows the correlation versus mean WAS for pairs of descriptors in the training set. In this figure, one of the descriptors is S.SIFT.1k. This (Pearson) correlation is computed based on the geographic distance of estimated location from the ground truth for each "test image" in the training set. The mean WAS is computed by $\frac{WAS(S.SIFT.1k)+WAS(X)}{2}$, where X is one of the other descriptors. As seen in this plot, the lowest correlation is for S.SIFT.1k and D.SIFT.1k, the latter being the third best mean WAS. We also observe that the correlation of S.SIFT.1k and each of the remaining descriptors are almost the same, but with different mean WAS.

Figure 4.27 shows the WAS results, in the training set, for the combination involving pairs of descriptors and one for the fusion of three descriptors: S.SIFT.1k, D.SIFT.1k (least correlated), and WAS.10k (best mean WAS). The last item in the figure presents the best result considering a descriptor alone to serve as baseline. Note that we used

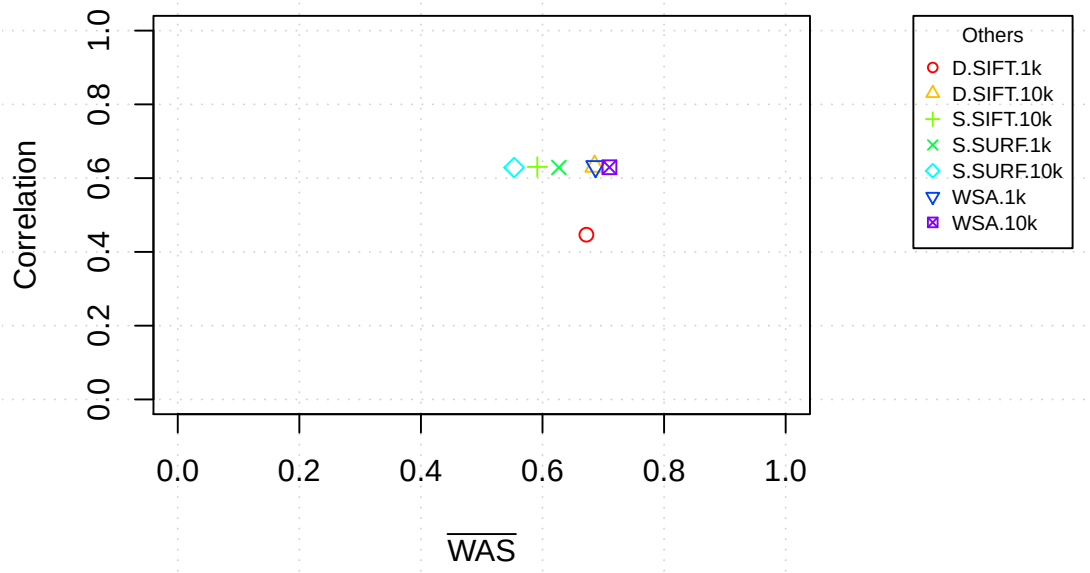
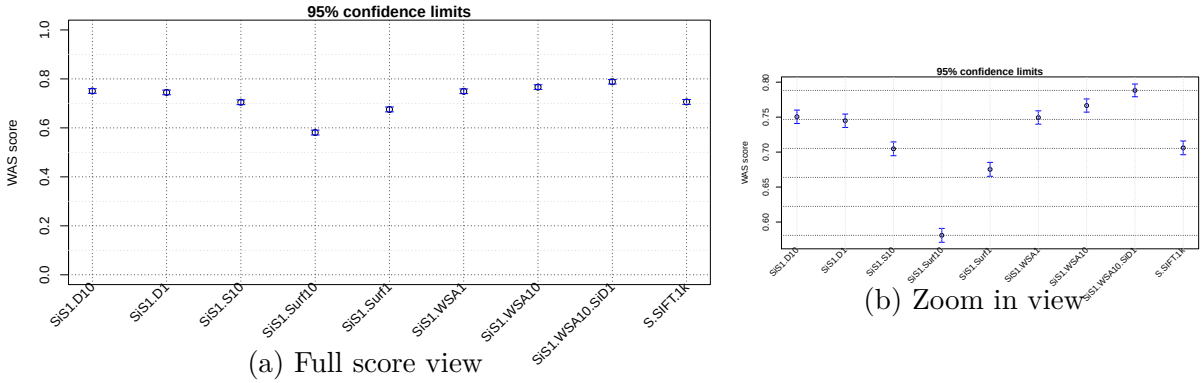


Figure 4.26: Correlation × mean WAS score between S.SIFT.1k and other descriptors in the *training* set.

the rank aggregation method defined in Equation 4.6. In the graph, we use a different (shorten) notation for naming the descriptors: SiS1 stands for S.SIFT.1K, D1 stands for the D.SIFT.1k, D10 stands for D.SIFT.10k, S10 stands for S.SIFT.10k, Surf1 stands for S.SURF.1k, Surf10 stands for S.SURF.10k, WSA1 stands for WSA.1k, and WSA10 stands for WSA.10k. For example, SiS1.WSA10 refers to the fusion result combining S.SIFT.1K and WSA.10k. As we can observe, the WAS resulting from the combination of two descriptors is close to the mean of their individual WAS. The highest result is yielded by the SiS1.WSA10 combination. Although this result is not significantly better than those observed for other combinations, such as SiS1.WSA1, SiS1.D1, and SiS1.D10, they are significantly better than the best single result (S.SIFT.1K). Note also that the fusion of three descriptors (SiS1.WSA10.D1) did significantly improve the result (0.7882) over those fusion pairs.

Test set Results and Discussion

In this section, we present and discuss data fusion results for the test set. In Figure 4.21 we presented the geocoding results for evaluated methods considering different precision levels in accumulative count of correctly assigned test images. The result exhibited in Figure 4.28 use the WAS score. As stated before, S.SURF.10k yielded the worst result for the training set, but the best in the test set. In fact, we can see that the other methods had lowered their results in the test set whereas S.SURF.10k maintained its performance with

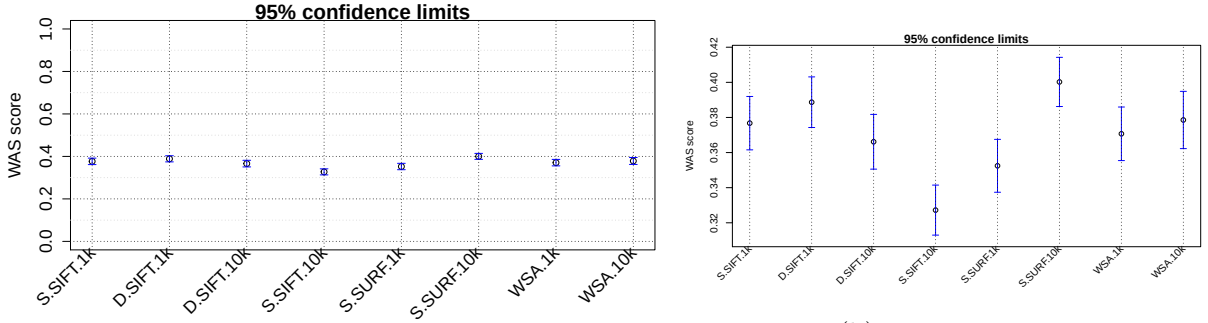


(a) Full score view

(b) Zoom in view

Figure 4.27: WAS scores and its confidence intervals for fusion results in the *training* set.

WAS score around 0.4 in both data sets. Additionally, we should note that S.SURF.10k is not statistically better than D.SIFT.1K, S.SIFT.1K, and WSA.10k.



(a) Full score view

(b) Zoom in view

Figure 4.28: WAS scores and its confidence intervals for single results in the *test* set.

Figure 4.29 shows the correlation analysis for the test set. Compared to the training set correlogram, the results on the test set also show a rather different behavior of the studied methods. In general, they seem to be less correlated (lighter color and less painted circles). We can also observe that S.SURF.10 has the lowest correlation coefficients.

Figure 4.30 shows the correlation versus mean WAS involving S.SIFT.1k and other descriptors. This figure shows a concentration of fusion methods for mean values in the range of 0.35 and 0.39. Note also that correlation scores are below 0.27.

The outcome of the fusion of S.SIFT.1k and S.Surf.10k was surprising. Unlike the other combinations, this fusion lowered the WAS. Moreover, looking at Figure 4.31, there is no clear winner in terms of best descriptor fusion, as most of the combinations yield statistically similar results. In addition to that, we also observe that the combination of SiS1, WSA10, and SiD1 does not yield enhanced results, as we had expected given the

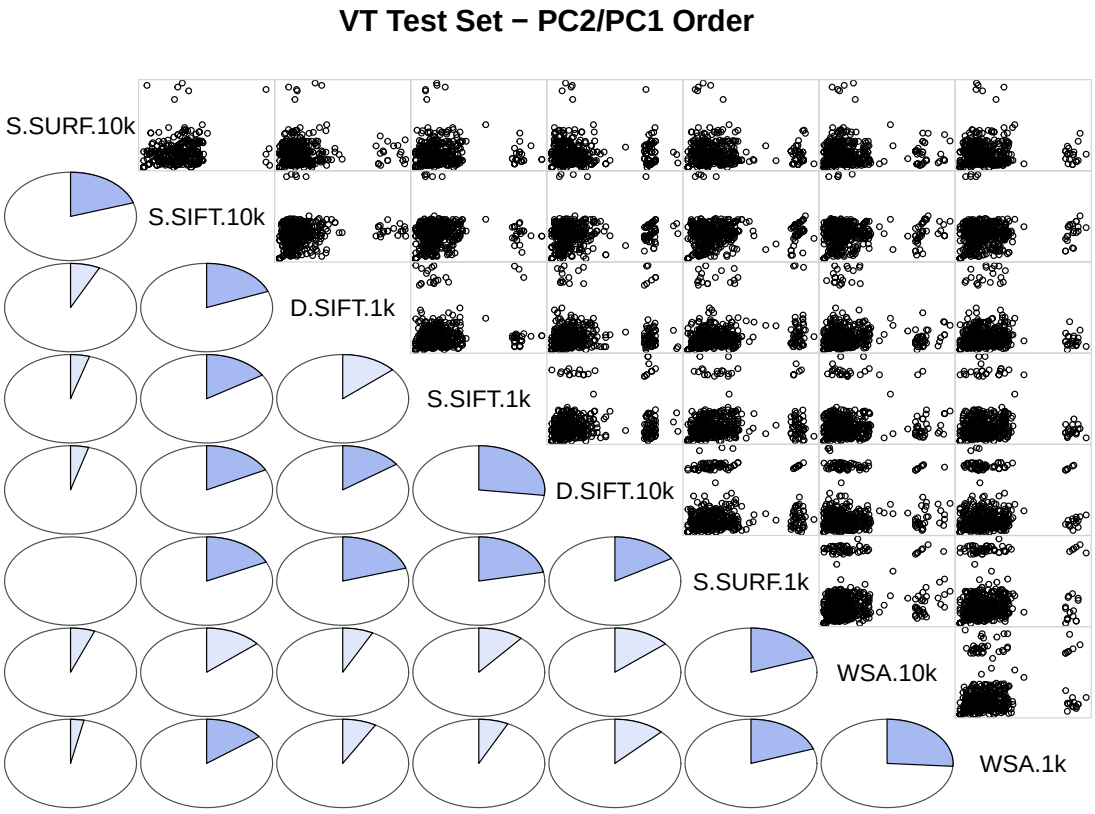


Figure 4.29: Correlation in the *test* set. Upper panel shows the dispersion graph for each pair of methods.

performance of this combination for the training set, nor the fusion improved over the best single descriptor result in test set (S.SURF.10k).

It seems that there are some fundamental differences between the training set and the test set that explains these distinct geocoding results of the evaluated methods. In fact, in the test set, there are more close up photos, whereas the training set includes pictures in a wider frame, i.e., the images depict more distant objects (buildings). These results suggest that a better performance of our geocoding framework in terms of combining different descriptors can be achieved when both the training and the test set share similar characteristics, as we observed in the Placing Task in the MediaEval 2012.

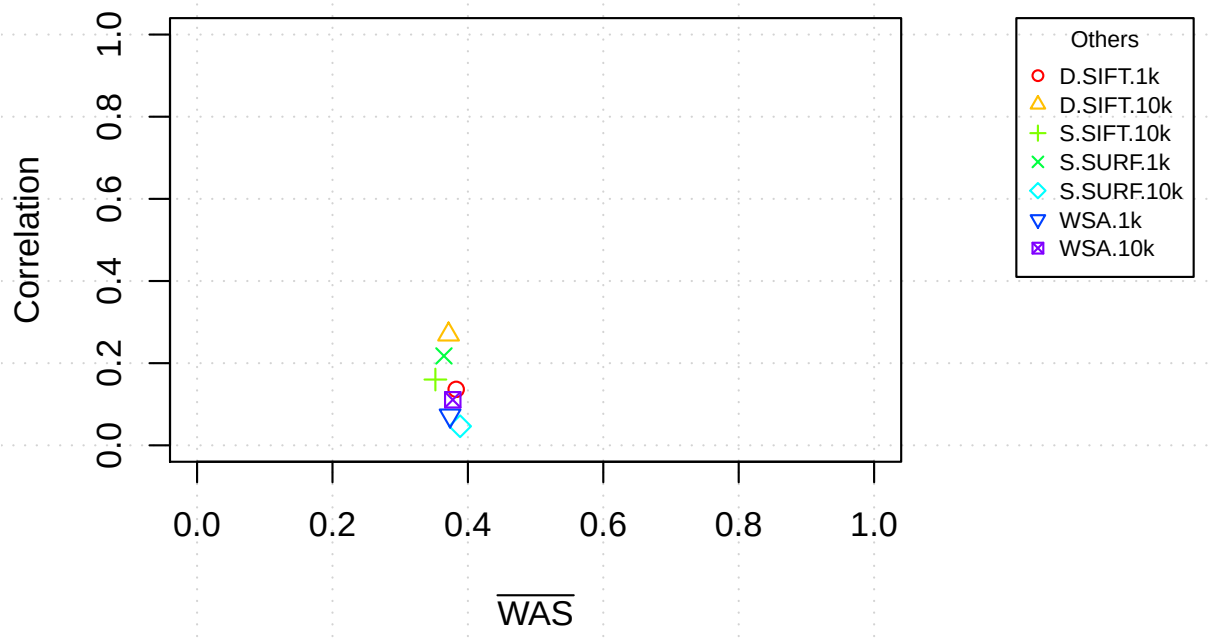
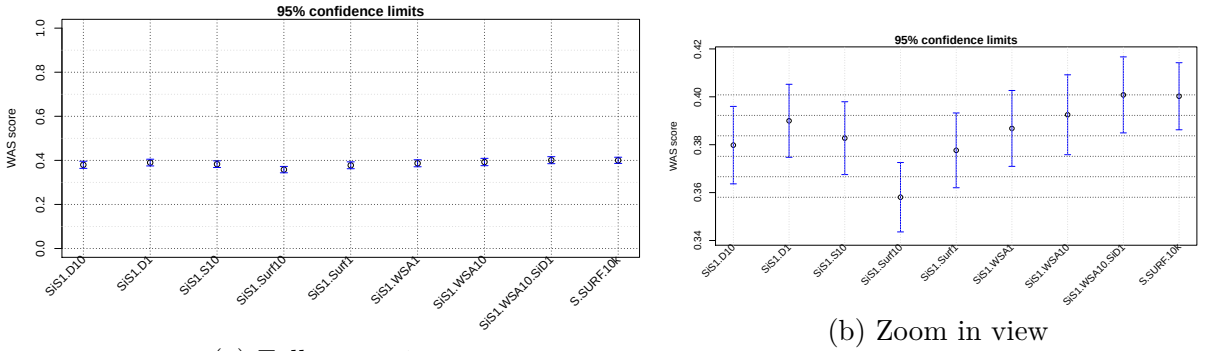


Figure 4.30: Correlation × mean WAS scores between S.SIFT.1k and other descriptors in the *test* set.



(a) Full score view

(b) Zoom in view

Figure 4.31: SSift1k × other WAS scores and its confidence interval for fusion results in *test* set.

Chapter 5

Conclusions

In this chapter, we summarize our contributions, as well as list the main conclusions of our study. We also present possible extensions to be addressed in future work.

5.1 Main contributions

References to places are often found in digital objects (e.g., documents, images, and videos) of several digital libraries. Geographic information can enrich services like browsing and searches, opening new opportunities for establishing new relations based on geographic location. However, the primary requirement to offer these geo-enabled services is that digital objects must be geocoded, i.e., they should be related to some place on Earth.

This work has investigated the fusion of textual and visual content for geocoding digital objects and has proposed a flexible framework to perform multimodal geocoding by combining ranked lists defined in terms of different modalities. In our approach, textual and visual descriptors were combined using a rank aggregation approach. To the best of our knowledge, this is the first attempt to address this problem using this kind of solution. The potential of this framework relies on the fact that each module can be improved separately, opening new opportunities for further investigation related to the development and use of novel rank aggregation methods for geocoding task. Moreover, it facilitates attaching new components to deal with new modalities of information and replacing the implementations by those that advance in dealing with a certain modality.

An architecture was designed to implement the proposed geocoding framework and it was first validated in the context of the Placing Task of the MediaEval 2012 initiative. Conducted experiments demonstrate that the use of the proposed fusion approach yields better results when compared with those based on a single clue (either textual or visual descriptor). Results also demonstrate that, despite of the simple textual descriptor methods used, the performance of the proposed method is comparable to the best submissions

of the Placing Task in the same year. We have also demonstrated the use of a method based on correlation analysis of geocoding results associated with different modalities to select features that enhanced geocoding results when combined.

In the context of photo geocoding of VT buildings, we have downloaded and selected the photos that could be used for the training dataset and produced the test set and prepared the ground truth location for them. In the VT building experiment, we were able to enhance significantly the geocoding results by using rank aggregation approaches if the training set (our knowledge base) captures the characteristics of the test set. The fusion framework was applied to combine results from different local visual descriptors.

Another contribution of this work is the proposal of a new score measure, named Weighted Average Score (WAS), to assess the quality of the results of geocoding methods. Instead of counting videos correctly assigned within various predefined precision radii (a common approach used in the literature), each method is evaluated in terms of a score between 0 (poor) and 1 (perfect), based on the geographical distances among produced predictions and ground truth locations. We showed that the proposed evaluation measure can be tuned to be used in specific experiment contexts (e.g., whole world or a city). This proposed measure was also implemented and applied in both study cases where we have validated our framework proposal.

5.2 Possible Extensions

We envision many further steps and derived works from this one. They are summarized below:

- Investigating other strategies for combining different modalities and exploring the strength of each modality for geocoding multimedia objects. Some promising alternatives rely on the use of rank aggregation methods based on re-ranking approaches [101, 102, 103, 104]. In addition, we would like to evaluate the use of supervised methods for feature selection and fusion as those applied in [31, 35].

- Using other external sources, such as user profiles and its indirect relationships, Geonames,[1] and Wikipedia [29], to filter out noisy data from ranked lists. Strategies used in [48, 122] can be combined with the proposed geocoding framework.

- Implementing novel components in the proposed framework that exploit ontologies, thesauri, and gazetteers to improve geocoding results.

[1] `http://www.geonames.org/` *(as of Dec. 2013).*

- Designing one additional level of abstraction in the fusion module by taking the top-k best ranked items found in ranked lists defined by different descriptors and use their geographical coordinates to combine or re-rank candidate locations. For example, we may consider the top-k points from a final rank aggregation list to define the location of a query object. Another strategy is to consider the coordinates of top-k items from ranked lists of different features. In these cases, we are addressing the geocoding problem by processing a set of candidate location points.

  We devise two possible methods following these ideas: (a) first, we can divide the Earth in a grid (e.g., fixed or varied size based on, for example, the approach proposed in [48]) and consider the cells where objects are more frequently found to reason a candidate location for the query image/video; (b) another strategy is to explore different clustering methods on candidate geographic coordinates and select the most promising cluster (e.g., the most dense) that will give its lat/long to the query object.

- Other possible extensions consist in the use of the proposed framework in different data fusion applications:

  - event detection tasks based on their type, time, and geographic location. In this case, the knowledge base should be about targeting events instead of locations;
  - geographic information retrieval of scholarly digital libraries based on image, sound, text, location, and time. For example, in [95], the fish species identification task is based on narrowing down the candidate family based on locations where a specimen was found besides other usual physical characteristics. Additionally, it might be interesting to explore the combination of the ranked lists from the searches of each individual feature in order to improve the final result;
  - social media geographic information mining based on multimodal data (e.g., image, text, time, users, friends and/or followers[28], preferences, and whereabouts);

- Incorporating other kinds of data such as those derived from remote sensing images (e.g., aerial), temporal information, and thematic maps (e.g., vegetation, terrain and population) into the proposed framework.

- Improving searching and browsing services by using the proposed framework to return more relevant documents. The proposed framework could be used for example to combine diverse features of a digital object (e.g., textual, visual, spatial-temporal). Geocoded digital objects can also be used in novel DL services such as search based on geographical queries as discussed in Section 2.1.2.

5.3 Published Contributions

This work resulted in publication of journal papers [82, 85], a book chapter [83], papers in conference proceedings [84, 86, 105], working notes [77, 78, 79], and technical reports [80, 81], which are listed below:

- **A rank aggregation framework for video multimodal geocoding** [85]. Lin Tzy Li, Daniel C. G. Pedronette, Jurandy Almeida, Otávio A. B. Penatti, Rodrigo Tripodi Calumby, and Ricardo da Silva Torres. *Multimedia Tools and Applications*, pages 1–37, 2013. http://dx.doi.org/10.1007/s11042-013-1588-4.

- **Revisitando os desafios da recuperação de informação geográfica na web** [82]. Lin Tzy Li and Ricardo da Silva Torres. *Cadernos CPqD Tecnologia*, 6(1):7–20, jan–jun 2010. http://www.cpqd.com.br/cadernosdetecnologia/Vol6_N1_jan_jun_2010/pdf/artigo1.pdf

- **Geospatial information** [83]. Lin Tzy Li and Ricardo da S. Torres. In Edward A. Fox and Jonathan P. Leidig, editors, *Digital Library Applications: CBIR, Education, Social Networks, eScience/Simulation, GIS*, Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, pages 85–120, March 2014. http://dx.doi.org/10.2200/S00565ED1V01Y201401ICR032

- **A visual approach for video geocoding using bag-of-scenes** [105]. Otávio A. B. Penatti, Lin Tzy Li, Jurandy Almeida, and Ricardo da Silva Torres. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, ICMR'12, pages 53:1–53:8, 2012. http://doi.acm.org/10.1145/2324796.2324857.

- **Domain-specific image geocoding: a case study on Virginia Tech building photos.** [86]. Lin Tzy Li, Otávio A. B. Penatti, Edward A. Fox, and Ricardo da Silva Torres. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL'13, pages 363–366, 2013. http://doi.acm.org/10.1145/2467696.2467727

- **Multimedia multimodal geocoding** [84]. Lin Tzy Li, Daniel Carlos Guimarães Pedronette, Jurandy Almeida, Otávio A. B. Penatti, Rodrigo Tripodi Calumby, and Ricardo da Silva Torres. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'12, pages 474–477, 2012. http://doi.acm.org/10.1145/2424321.2424393

- **CTRnet DL for disaster information services** [129]. Seungwon Yang, Andrea L. Kavanaugh, Nádia P. Kozievitch, Lin Tzy Li, Venkat Srinivasan, Steven D.

Sheetz, Travis Whalen, Donald Shoemaker, Ricardo da Silva Torres, and Edward A. Fox. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, JCDL'11, pages 437–438, 2011. http://doi.acm.org/10.1145/1998076.1998173

- **RECOD working notes for placing task MediaEval 2011** [77]. Lin Tzy Li, Jurandy Almeida, and Ricardo da Silva Torres. In *Working Notes Proceedings of the MediaEval Workshop*, volume 807 of *CEUR Workshop Proceedings*, pages 1–2, 2011. http://ceur-ws.org/Vol-807/Li_UNICAMP_Placing_me11wn.pdf

- **A multimodal approach for video geocoding** [78]. Lin Tzy Li, Jurandy Almeida, Daniel C. G. Pedronette, Otávio A. B. Penatti, and Ricardo da Silva Torres. In *Working Notes Proceedings of the MediaEval 2012 Workshop*, volume 927 of *CEUR Workshop Proceedings*, 2 pages, 2012. http://ceur-ws.org/Vol-927/mediaeval2012_submission_19.pdf

- **Multimodal image geocoding: the 2013 RECOD's approach** [79]. Lin Tzy Li, Jurandy Almeida, Otávio A. B. Penatti, Rodrigo Tripodi Calumby, Daniel André Gonçalves, and Ricardo da Silva Torres. In *Working Notes Proceedings of the MediaEval Workshop*, volume 1043 of *CEUR Workshop Proceedings*, pages 1–2, 2013. http://ceur-ws.org/Vol-1043/mediaeval2013_submission_65.pdf

- **Coping with geographical relationships in web searches** [81]. Lin Tzy Li and Ricardo da Silva Torres. Technical Report IC-10-04, Institute of Computing, University of Campinas, 19 pages, January 2010. http://www.ic.unicamp.br/~reltech/2010/10-04.pdf

- **Revisitando os desafios da recuperação de informação geográfica na web** [80]. Lin Tzy Li and Ricardo da Silva Torres. Technical Report IC-09-18, Institute of Computing, University of Campinas, 19 pages, May 2009. http://www.ic.unicamp.br/~reltech/2009/09-18.pdf

We also collaborated on some initiatives that have been carried out in Digital Library Research Lab (DLRL) at Virginia Tech, USA, in the context of two research projects. Those efforts resulted in co-authoring the following additional works:

- **SuperIDR: a tool for fish identification and information retrieval** [94]. Uma Murthy, Edward A. Fox, Yinlin Chen, Eric M. Hallerman, Donald J. Orth, Ricardo da Silva Torres, Lin Tzy Li, Nádia P. Kozievitch, Felipe S. P. Andrade, Tiago R. C. Falcão, and Evandro Ramos. *Fisheries*, 38(2):65–75, 2013. http://www.tandfonline.com/doi/abs/10.1080/03632415.2013.757982

- **Social media use by government: from the routine to the critical** [58]. Andrea L. Kavanaugh, Edward A. Fox, Steven D. Sheetz, Seungwon Yang, Lin Tzy Li, Donald J. Shoemaker, Apostol Natsev, and Lexing Xie. *Government Information Quarterly*, 29(4):480 – 491, 2012. Social Media in Government - Selections from the 12th Annual International Conference on Digital Government Research (dg.o2011). http://dx.doi.org/10.1016/j.giq.2012.06.002

- **Social media use by government: from the routine to the critical**[59]. Andrea Kavanaugh, Edward A. Fox, Steven D. Sheetz, Seungwon Yang, Lin Tzy Li, Travis Whalen, Donald Shoemaker, Paul Natsev, and Lexing Xie. In *Proceedings of the International Digital Government Research Conference*, dg.o '11, 10 pages, 2011. http://doi.acm.org/10.1145/2037556.2037574

- **Use of subimages in fish species identification: a qualitative study** [95]. Uma Murthy, Lin Tzy Li, Eric Hallerman, Edward A. Fox, Manuel A. Pérez-Quiñones, Lois M. Delcambre, and Ricardo da Silva Torres. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '11, pages 185–194, 2011. http://doi.acm.org/10.1145/1998076.1998112

- **Experiment and analysis services in a fingerprint digital library for collaborative research** [98]. Sung Hee Park, Jonathan P. Leidig, Lin Tzy Li, Edward A. Fox, Nathan J. Short, Kevin E. Hoyle, A. Lynn Abbott, and Michael S. Hsiao. In *Research and Advanced Technology for Digital Libraries. Proceedings of International Conference on Theory and Practice of Digital Libraries*, TPDL 2011, volume 6966 of *Lecture Notes in Computer Science*, pages 179–191, 2011. http://doi.acm.org/10.1145/2042536.2042562

- **Microblogging in crisis situations: Mass protests in Iran, Tunisia, Egypt** [60]. Andrea Kavanaugh, Seungwon Yang, Steve Sheetz, Lin Tzy Li, and Edward A. Fox. In *TRANSNATIONAL HCI Workshop in conjuction with the ACM Conference on Human Factors in Computing Systems (CHI'11)*, 7 pages, 2011. http://www.princeton.edu/~jvertesi/TransnationalHCI/Participants_files/Kavanaugh.pdf

- **Twitter use during an emergency event: the case of the UT Austin shooting** [87]. Lin Tzy Li, Seungwon Yang, Andrea Kavanaugh, Edward A. Fox, Steven D. Sheetz, Donald Shoemaker, Travis F. Whalen, and Venkat Srinivasan. In *Proceedings of the International Digital Government Research Conference*, dg.o '11, 2 pages, 2011. http://doi.acm.org/10.1145/2037556.2037613

Bibliography

[1] Mirna Adriani and Monica Lestari Paramita. Identifying location in Indonesian documents for geographic information retrieval. In *Proceedings of the ACM Workshop on Geographic Information Retrieval*, GIR'07, pages 19–24, 2007.

[2] Dirk Ahlers and Susanne Boll. Retrieving address-based locations from the web. In *Proceedings of the ACM Workshop on Geographic Information Retrieval*, GIR'08, pages 27–34, 2008.

[3] Jurandy Almeida, Neucimar Jerônimo Leite, and Ricarda da Silva Torres. Comparison of video sequences with histograms of motion patterns. In *Proceedings of the International Conference on Image Processing*, ICIP'11, pages 3673–3676, 2011.

[4] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. Web-a-where: Geotagging web content. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'04, pages 273–280, 2004.

[5] David S. Batista, Mário J. Silva, Francisco M. Couto, and Bibek Behera. Geographic signatures for semantic retrieval. In *Proceedings of the ACM Workshop on Geographic Information Retrieval*, GIR'10, pages 19:1–19:8, 2010.

[6] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision – ECCV 2006. Proceedings of the European Conference on Computer Vision (Part I)*, volume 3951 of *Lecture Notes in Computer Science*, pages 404–417, 2006.

[7] Andreas D. Blaser and Max J. Egenhofer. A visual tool for querying geographic databases. In *Proceedings of the Working Conference on Advanced visual interfaces*, AVI'00, pages 211–216, 2000.

[8] Andre Blessing, Reinhard Kuntz, and Hinrich Schütze. Towards a context model driven German geo-tagging system. In *Proceedings of the ACM Workshop on Geographic Information Retrieval*, GIR'07, pages 25–30, 2007.

[9] Karla A. V. Borges. *Uso de uma ontologia de lugar urbano para reconhecimento e extração de evidências geoespaciais na Web*. Doctoral thesis, UFMG - Universidade Federal de Minas Gerais, 2006.

[10] Karla A. V. Borges, Clodoveu A. Davis, Alberto H. F. Laender, and Claudia B. Medeiros. Ontology-driven discovery of geospatial evidence in web pages. *GeoInformatica*, 15(4):609–631, 2011.

[11] Karla A. V. Borges, Alberto H. F. Laender, Claudia B. Medeiros, and Clodoveu A. Davis Jr. Discovering geographic locations in web pages using urban addresses. In *Proceedings of the ACM Workshop on Geographic Information Retrieval*, GIR'07, pages 31–36, 2007.

[12] Y-Lan Boureau, Francis Bach, Yann LeCun, and Jean Ponce. Learning mid-level features for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2010, pages 2559–2566, 2010.

[13] Daniela F. Brauner, Marco A. Casanova, and Ruy L. Milidiü. Towards gazetteer integration through an instance-based thesauri mapping approach. In *Advances in Geoinformatics. Brazilian Symposium on GeoInformatics*, GEOINFO'06, pages 235–245, Campos do Jordão, SP, Brazil, 2007.

[14] Davide Buscaldi and Paulo Rosso. A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science*, 22(3):301, 2008.

[15] Rodrigo Tripodi Calumby. Recuperação multimodal de imagens com realimentação de relevância baseada em programação genética. Master's thesis, Universidade Estadual de Campinas (UNICAMP), Instituto de Computação, Campinas, SP, Brazil, 2010.

[16] Rodrigo Tripodi Calumby, Ricardo da Silva Torres, and Marcos André Gonçalves. Multimodal retrieval with relevance feedback based on genetic programming. *Multimedia Tools and Applications*, 69(3):991–1019, 2014.

[17] Gilberto Câmara, Marco A. Casanova, Andréa S. Hemerly, Geovane C. Magalhães, and Cláudia M. B. Medeiros. Anatomia de sistemas de informação geográfica. In *10a. Escola de Computação*, page 197, Campinas, 1996. Instituto de Computação - UNICAMP.

[18] Cláudio E. C. Campelo and Cláudio de Souza Baptista. Geographic scope modeling for web documents. In *Proceedings of the ACM Workshop on Geographic Information Retrieval*, GIR'08, pages 11–18, 2008.

[19] Nuno Cardoso and Mário J. Silva. Query expansion through geographical feature types. In *Proceedings of the ACM Workshop on Geographic Information Retrieval*, GIR'07, pages 55–60, 2007.

[20] Yen-Yu Chen, Torsten Suel, and Alexander Markowetz. Efficient query processing in geographic web search engines. In *Proceedings of ACM SIGMOD international conference on Management of data*, SIGMOD'06, pages 277–288, 2006.

[21] Jaeyoung Choi, Venkatesan N. Ekambaram, Gerald Friedland, and Kannan Ramchandran. The 2012 ICSI/Berkeley video location estimation system. In Larson et al. [73], pages 1–2.

[22] Jaeyoung Choi, Howard Lei, and Gerald Friedland. The 2011 ICSI video location estimation system. In *Working Notes Proceedings of the MediaEval Workshop*, volume 807 of *CEUR Workshop Proceedings*, pages 1–2, 2011.

[23] Eliseo Clementini, Paolino Di Felice, and Peter van Oosterom. A small set of formal topological relationships suitable for end-user interaction. In *Proceedings of the International Symposium on Advances in Spatial Databases*, SSD'91, pages 277–295, 1993.

[24] Stephane Clinchant, Julien Ah-Pine, and Gabriela Csurka. Semantic combination of textual and visual information in multimedia retrieval. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, ICMR'11, pages 44:1–44:8, 2011.

[25] Don Coppersmith, Lisa K. Fleischer, and Atri Rurda. Ordering by weighted number of wins gives a good ranking for weighted tournaments. *ACM Transactions on Algorithms*, 6(3):55:1–55:13, July 2010.

[26] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'09, pages 758–759, 2009.

[27] W. Bruce Croft. Combining approaches to information retrieval. In W. Bruce Croft, editor, *Advances in Information Retrieval. Recent Research from the Center for Intelligent Information Retrieval*, volume 7 of *The Information Retrieval*, pages 1–36. Springer US, 2000.

[28] Clodoveu A. Davis Jr., Gisele L. Pappa, Diogo Rennó Rocha de Oliveira, and Filipe de L. Arcanjo. Inferring the location of twitter messages based on user relationships. *Transactions in GIS*, 15(6):735–751, 2011.

[29] Rafael Odon de Alencar, Clodoveu Augusto Davis Jr., and Marcos André Gonçalves. Geographical classification of documents using evidence from wikipedia. In *Proceedings of the ACM Workshop on Geographic Information Retrieval*, GIR'10, pages 12:1–12:8, 2010.

[30] Dayong Ding and Bo Zhang. Probabilistic model supported rank aggregation for the semantic concept detection in video. In *Proceedings of the International Conference on Image and Video Retrieval*, CIVR'07, pages 587–594, 2007.

[31] Jefersson Alex dos Santos, Philippe-Henri Gosselin, Sylvie Philipp-Foliguet, Ricardo da Silva Torres, and Alexandre X. Falcão. Multiscale classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 50(10):3764–3775, 2012.

[32] Max J. Egenhofer. Query processing in spatial-query-by-sketch. *Journal of Visual Languages & Computing*, 8:403–424, August 1997.

[33] Weiguo Fan, Edward A. Fox, Praveen Pathak, and Harris Wu. The effects of fitness functions on genetic programming-based ranking discovery for web search. *Journal of the American Society for Information Science and Technology*, 55(7):628–636, 2004.

[34] Weiguo Fan, Praveen Pathak, and Linda Wallace. Nonlinear ranking function representations in genetic programming-based ranking discovery for personalized search. *Decision Support Systems*, 42(3):1338 – 1349, 2006.

[35] Fabio Augusto Faria, Adriano Veloso, Humberto Mossri de Almeida, Eduardo do Valle, Ricardo da Silva Torres, Marcos André Gonçalves, and Wagner Meira Jr. Learning to rank for content-based image retrieval. In *Proceedings of the International Conference on Multimedia Information Retrieval*, MIR'10, pages 285–294, 2010.

[36] Damires Fernandes and Ana Carolina Salgado. Geovisual interface - a visual query interface for geographic information systems. In *Proceedings of the Brazilian Symposium on Databases*, SBBD'00, pages 7–19, 2000.

[37] Peter C. Fishburn. *Nonlinear preference and utility theory / Peter C. Fishburn.* Johns Hopkins University Press Baltimore, 1988.

[38] Edward A. Fox, Christopher Andrews, Weiguo Fan, Jian Jiao, Ananya Kassahun, Szu-Chia Lu, Yifei Ma, Chris North, Naren Ramakrishnan, Angela Scarpa, Bruce H.

Friedman, and Steven D. Sheetz. A Digital Library for Recovery, Research, and Learning From April 16, 2007, at Virginia Tech. *Traumatology*, 14(1), 2008.

[39] Edward A. Fox and Joseph A. Shaw. Combination of multiple searches. In *Proceedings of the Text REtrieval Conference (TREC)*, volume 500-215 of *NIST Special Publication, TREC-2*, pages 243–252, 1994.

[40] Michael Freeston. The Alexandria digital library and the Alexandria digital earth prototype. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL'04, pages 410–410, 2004.

[41] James Frew, Michael Freeston, Nathan Freitas, Linda Hill, Greg Janee, Kevin Lovette, Robert Nideffer, Terence Smith, and Qi Zheng. The Alexandria digital library architecture. *International Journal on Digital Libraries*, 2(4):259–268, May 2000.

[42] Michael Friendly. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56(4):316–324, January 2002.

[43] Gaihua Fu, Christopher B. Jones, and Alia I. Abdelmoty. Ontology-based spatial query expansion in information retrieval. In *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE. Proceedings of OTM Confederated International Conferences, CoopIS, DOA, and ODBASE*, volume 3761 of *Lecture Notes in Computer Science*, pages 1466–1482. Springer Berlin Heidelberg, 2005.

[44] A. Gallagher, D. Joshi, Jie Yu, and Jiebo Luo. Geo-location inference from image content and user tags. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, CVPR 2009, pages 55–62. IEEE, June 2009.

[45] Fatiha Guerroudji-Meddah, Hafida Belbachir, and Robert Laurini. A visual language for GIS querying. In *IEEE International Conference on Computer Science and Information Technology*, ICCSIT'2009, pages 518–521, August 2009.

[46] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, 2 edition, 2009. `http://statweb.stanford.edu/~tibs/ElemStatLearn/`.

[47] Claudia Hauff and Geert-Jan Houben. WISTUD at MediaEval 2011: Placing task. In *Working Notes Proceedings of the MediaEval Workshop*, volume 807 of *CEUR Workshop Proceedings*, pages 1–2, 2011.

[48] Claudia Hauff and Geert-Jan Houben. Placing images on the world map: A microblog-based enrichment approach. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'12, pages 691–700, 2012.

[49] James Hays and Alexei A. Efros. IM2GPS: estimating geographic information from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2008, pages 1–8, 2008.

[50] Daqing He. A study of self-organizing map in interactive relevance feedback. In *Proceedings of the International Conference on Information Technology: New Generations*, ITNG'06, pages 394–401. IEEE, 2006.

[51] Andreas Henrich and Volker Luedecke. Characteristics of geographic information needs. In *Proceedings of the ACM Workshop on Geographic Information Retrieval*, GIR'07, pages 1–6, 2007.

[52] Greg Janée and James Frew. The ADEPT digital library architecture. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL'02, pages 342–350, 2002.

[53] Christopher B. Jones. Geographic information retrieval, 2006. Keynote Presentation at GEOINFO 2006. Campos do Jordão, SP, Brazil. http://www.geoinfo.info/geoinfo2006/presentation/Christopher_Jones.ppt.

[54] Christopher B. Jones, Alia I. Abdelmoty, David Finch, Gaihua Fu, and Subodh Vaid. The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing. In *Geographic Information Science. Proceedings of International Conference on Geographic Information Science*, volume 3234 of *Lecture Notes in Computer Science*, pages 125–139, 2004. GIScience 2004.

[55] Christopher B. Jones and Ross S. Purves. Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3):219–228, March 2008.

[56] Rosie Jones, Ahmed Hassan, and Fernando Diaz. Geographic features in web search retrieval. In *Proceedings of the ACM Workshop on Geographic Information Retrieval*, GIR'08, pages 57–58, 2008.

[57] Yannis Kalantidis, Giorgos Tolias, Yannis Avrithis, Marios Phinikettos, Evaggelos Spyrou, Phivos Mylonas, and Stefanos Kollias. VIRaL: Visual image retrieval and localization. *Multimedia Tools and Applications*, 51(2):555–592, January 2011.

[58] Andrea L. Kavanaugh, Edward A. Fox, Steven D. Sheetz, Seungwon Yang, Lin Tzy Li, Donald J. Shoemaker, Apostol Natsev, and Lexing Xie. Social media use by government: from the routine to the critical. *Government Information Quarterly*, 29(4):480–491, 2012. Social Media in Government - Selections from dg.o2011.

[59] Andrea L. Kavanaugh, Edward A. Fox, Steven D. Sheetz, Seungwon Yang, Lin Tzy Li, Travis Whalen, Donald Shoemaker, Paul Natsev, and Lexing Xie. Social media use by government: from the routine to the critical. In *Proceedings of the International Digital Government Research Conference*, dg.o'11, pages 121–130, 2011.

[60] Andrea L. Kavanaugh, Seungwon Yang, Steve Sheetz, Lin Tzy Li, and Edward A. Fox. Microblogging in Crisis Situations: Mass Protests in Iran, Tunisia, Egypt. In *TRANSNATIONAL HCI Workshop in conjuction with the ACM Conference on Human Factors in Computing Systems (CHI'11)*, pages 1–7, 2011.

[61] Pascal Kelm, Sebastian Schmiedeke, and Thomas Sikora. A hierarchical, multi-modal approach for placing videos on the map using millions of flickr photographs. In *Workshop on Social and Behavioural Networked Media Access*, SBNMA'11, pages 15–20, 2011.

[62] Pascal Kelm, Sebastian Schmiedeke, and Thomas Sikora. Multi-modal, Multi-resource Methods for Placing Flickr Videos on the Map. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, ICMR'11, pages 52:1–52:8, 2011.

[63] Pascal Kelm, Sebastian Schmiedeke, and Thomas Sikora. How spatial segmentation improves the multimodal geo-tagging. In Larson et al. [73], pages 1–2.

[64] Pascal Kelm, Sebastian Schmiedeke, and Thomas Sikora. Multimodal geo-tagging in social media websites using hierarchical spatial segmentation. In *Proceedings of the ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN'12, pages 32–39, 2012.

[65] Syed Awase Khirni, Bisheng Yang, Ross Purves, and Matthias Kopczynski. Query interface design. Project Report W4 D74101, University of Zurich, Zurich, Switzerland, 2003.

[66] Anna Khudyak Kozorovitsky and Oren Kurland. Cluster-based fusion of retrieved lists. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'11, pages 893–902, 2011.

[67] Alexandre Klementiev, Dan Roth, and Kevin Small. A framework for unsupervised rank aggregation. In *Proceedings of the ACM SIGIR Conference Workshop on Learning to Rank for Information Retrieval*, LR4IR 2008, pages 32–39, 7 2008.

[68] Jana Kludas, Eric Bruno, and Stéphane Marchand-Maillet. Information fusion in multimedia information retrieval. In Nozha Boujemaa, Marcin Detyniecki, and Andreas Nürnberger, editors, *Adaptive Multimedia Retrieval: Retrieval, User, and Semantics. Proceedings of International Workshop on Adaptive Multimedial Retrieval: Retrieval, User, and Semantics*, volume 4918 of *Lecture Notes in Computer Science*, pages 147–159, 2008. AMR 2007.

[69] Mieczyslaw M. Kokar, Jerzy A. Tomasik, and Jerzy Weyman. Formalizing classes of information fusion systems. *Information Fusion*, 5(3):189 – 202, 2004.

[70] Olivier Van Laere, Steven Schockaert, and Bart Dhoedt. Ghent university at the 2011 placing task. In *Working Notes Proceedings of the MediaEval Workshop*, volume 807, pages 1–2. CEUR-WS.org, 2011.

[71] Olivier Van Laere, Steven Schockaert, Jonathan A. Quinn, Frank C. Langbein, and Bart Dhoedt. Ghent and Cardiff university at the 2012 placing task. In Larson et al. [73], pages 1–2.

[72] Martha Larson, Mohammad Soleymani, Pavel Serdyukov, Stevan Rudinac, Christian Wartena, Vanessa Murdock, Gerald Friedland, Roeland Ordelman, and Gareth J. F. Jones. Automatic tagging and geotagging in video collections and communities. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, ICMR'11, pages 51:1–51:8, 2011.

[73] Martha A. Larson, Sebastian Schmiedeke, Pascal Kelm, Adam Rae, Vasileios Mezaris, Tomas Piatrik, Mohammad Soleymani, Florian Metze, and Gareth J. F. Jones, editors. *Working Notes Proceedings of the MediaEval 2012 Workshop*, volume 927 of *CEUR Workshop Proceedings*, 2012.

[74] Ray R. Larson. Placing cultural events and documents in space and time. In Matt Duckham, Michael F. Goodchild, and Michael Worboys, editors, *Foundations of Geographic Information Science*, pages 223–239. Taylor & Francis, 2003.

[75] Ray R. Larson. Geographic information retrieval and digital libraries. In *Research and Advanced Technology for Digital Libraries. Proceedings of European Conference on Digital Libraries*, volume 5714 of *Lecture Notes in Computer Science*, pages 461–464, 2009. ECDL 2009.

[76] Johannes Leveling and Sven Hartrumpf. On metonymy recognition for geographic information retrieval. *International Journal of Geographical Information Science*, 22(3):289, 2008.

[77] Lin Tzy Li, Jurandy Almeida, and Ricardo da Silva Torres. RECOD working notes for placing task MediaEval 2011. In *Working Notes Proceedings of the MediaEval Workshop*, volume 807 of *CEUR Workshop Proceedings*, pages 1–2, 2011.

[78] Lin Tzy Li, Jurandy Almeida, Daniel C. G. Pedronette, Otávio A. B. Penatti, and Ricardo da Silva Torres. A multimodal approach for video geocoding. In Larson et al. [73], pages 1–2.

[79] Lin Tzy Li, Jurandy Almeida, Otávio A. B. Penatti, Rodrigo Tripodi Calumby, Daniel C. G. Pedronette, Marcos André Gonçalves, and Ricardo da Silva Torres. Multimodal image geocoding: the 2013 RECOD's approach. In *Working Notes Proceedings of the MediaEval Workshop*, volume 1043 of *CEUR Workshop Proceedings*, pages 1–2, 2013.

[80] Lin Tzy Li and Ricardo da Silva Torres. Revisitando os desafios da recuperação de informação geográfica na web. Technical Report IC-09-18, Institute of Computing, University of Campinas, May 2009.

[81] Lin Tzy Li and Ricardo da Silva Torres. Coping with geographical relationships in web searches. Technical Report IC-10-04, Institute of Computing, University of Campinas, January 2010.

[82] Lin Tzy Li and Ricardo da Silva Torres. Revisitando os desafios da recuperação de informação geográfica na web. *Cadernos CPqD Tecnologia*, 6(1):7–20, jan–jun 2010.

[83] Lin Tzy Li and Ricardo da Silva Torres. Geospatial information. In Edward A. Fox and Jonathan P. Leidig, editors, *Digital Library Applications: CBIR, Education, Social Networks, eScience/Simulation, GIS*, Synthesis Lectures on Information Concepts, Retrieval, and Services, pages 85–120. Morgan & Claypool Publishers, San Francisco, March 2014.

[84] Lin Tzy Li, Daniel C. G. Pedronette, Jurandy Almeida, Otávio A. B. Penatti, Rodrigo Tripodi Calumby, and Ricardo da Silva Torres. Multimedia multimodal geocoding. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'12, pages 474–477, 2012.

[85] Lin Tzy Li, Daniel C. G. Pedronette, Jurandy Almeida, Otávio A. B. Penatti, Rodrigo Tripodi Calumby, and Ricardo da Silva Torres. A rank aggregation framework for video multimodal geocoding. *Multimedia Tools and Applications*, pages 1–37, 2013. http://dx.doi.org/10.1007/s11042-013-1588-4.

[86] Lin Tzy Li, Otávio A. B. Penatti, Edward A. Fox, and Ricardo da Silva Torres. Domain-specific image geocoding: a case study on Virginia Tech building photos. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL'13, pages 363–366, 2013.

[87] Lin Tzy Li, Seungwon Yang, Andrea L. Kavanaugh, Edward A. Fox, Steven D. Sheetz, Donald Shoemaker, Travis F. Whalen, and Venkat Srinivasan. Twitter use during an emergency event: the case of the UT Austin shooting. In *Proceedings of the International Digital Government Research Conference*, dg.o'11, pages 335–336, 2011.

[88] Xinchao Li, Claudia Hauff, Martha Larson, and Alan Hanjalic. Preliminary exploration of the use of geographical information for content-based geo-tagging of social video. In Larson et al. [73], pages 1–2.

[89] Jiebo Luo, Dhiraj Joshi, Jie Yu, and Andrew Gallagher. Geotagging in multimedia and computer vision–a survey. *Multimedia Tools and Applications*, 51(1):187–211, January 2011.

[90] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[91] Bruno Martins, Mário J. Silva, and Leonardo Andrade. Indexing and ranking in Geo-IR systems. In *Proceedings of the ACM Workshop on Geographic Information Retrieval*, GIR'05, pages 31–34, 2005.

[92] Mode. *Oxford Dictionaries*. Oxford Univerity Press, Mar 2014. `http://www.oxforddictionaries.com/us/definition/american_english/mode?q=mode`.

[93] Mark Montague and Javed A. Aslam. Condorcet fusion for improved retrieval. In *Proceedings of the International Conference on Information and Knowledge Management*, CIKM'02, pages 538–548, 2002.

[94] Uma Murthy, Edward A. Fox, Yinlin Chen, Eric M. Hallerman, Donald J. Orth, Ricardo da Silva Torres, Lin Tzy Li, Nádia P. Kozievitch, Felipe S. P. Andrade, Tiago R. C. Falcão, and Evandro Ramos. SuperIDR: a tool for fish identification and information retrieval. *Fisheries*, 38(2):65–75, 2013.

[95] Uma Murthy, Lin Tzy Li, Eric Hallerman, Edward A. Fox, Manuel A. Pérez-Quiñones, Lois M. Delcambre, and Ricardo da Silva Torres. Use of subimages in fish species identification: a qualitative study. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL'11, pages 185–194, 2011.

[96] Andreas M. Olligschlaeger and Alexander G. Hauptmann. Multimodal Information Systems and GIS: The Informedia Digital Video Library. In *1999 ESRI User Conference*, 1999.

[97] Simon Overell and Stefan Rüger. Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science*, 22(3):265, 2008.

[98] Sung Hee Park, Jonathan P. Leidig, Lin Tzy Li, Edward A. Fox, Nathan J. Short, Kevin E. Hoyle, A. Lynn Abbott, and Michael S. Hsiao. Experiment and analysis services in a fingerprint digital library for collaborative research. In *Research and Advanced Technology for Digital Libraries. Proceedings of International Conference on Theory and Practice of Digital Libraries*, volume 6966 of *Lecture Notes in Computer Science*, pages 179–191, 2011. TPDL 2011.

[99] Robert C. Pasley, Paul D. Clough, and Mark Sanderson. Geo-tagging for imprecise regions of different sizes. In *Proceedings of the ACM Workshop on Geographic Information Retrieval*, GIR'07, pages 77–82, 2007.

[100] Daniel C. G. Pedronette. *Exploiting Contextual Information for Image Re-Ranking and Rank Aggregation in Image Retrieval Tasks*. PhD thesis, University of Campinas (UNICAMP), Campinas, SP, Brazil, 2012.

[101] Daniel C. G. Pedronette and Ricardo da S. Torres. Image re-ranking and rank aggregation based on similarity of ranked lists. *Pattern Recognition*, 46(8):2350 – 2360, 2013.

[102] Daniel C. G. Pedronette and Ricardo da Silva Torres. Exploiting clustering approaches for image re-ranking. *Journal of Visual Languages & Computing*, 22(6):453–466, 2011.

[103] Daniel C. G. Pedronette, Ricardo da Silva Torres, and Rodrigo Tripodi Calumby. Using contextual spaces for image re-ranking and rank aggregation. *Multimedia Tools and Applications*, 69(3):689–716, 2014.

[104] Daniel C. G. Pedronette, Otávio A. B. Penatti, and Ricardo da S. Torres. Unsupervised manifold learning using reciprocal knn graphs in image re-ranking and rank aggregation tasks. *Image and Vision Computing*, 32(2):120–130, 2014.

[105] Otávio A. B. Penatti, Lin Tzy Li, Jurandy Almeida, and Ricardo da Silva Torres. A visual approach for video geocoding using bag-of-scenes. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, ICMR'12, pages 53:1–53:8, 2012.

[106] Otávio A. B. Penatti, Fernanda B. Silva, Eduardo Valle, Valerie Gouet-Brunet, and Ricardo da Silva Torres. Visual word spatial arrangement for image retrieval and classification. *Pattern Recognition*, 47(2):705–720, 2014.

[107] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2007, pages 1–8, 2007.

[108] Norman Poh and Samy Bengio. How do correlation and variance of base-experts affect fusion in biometric authentication tasks? *IEEE Transactions on Signal Processing*, 53(11):4384–4396, November 2005.

[109] Adrian Popescu and Nicolas Ballas. CEA LIST's participation at MediaEval 2012 placing task. In Larson et al. [73], pages 1–2.

[110] Adrian Popescu, Gregory Grefenstette, and Pierre Alain Moëllic. Gazetiki: automatic creation of a geographical gazetteer. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL'08, pages 85–93, 2008.

[111] Ross S. Purves, Paul Clough, Christopher B. Jones, Avi Arampatzis, Benedicte Bucher, David Finch, Gaihua Fu, Hideo Joho, Awase Khirni Syed, Subodh Vaid, and Bisheng Yang. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the internet. *International Journal of Geographical Information Science*, 21(7):717–745, 2007.

[112] Adam Rae and Pascal Kelm. Working notes for the placing task at MediaEval 2012. In Larson et al. [73], pages 1–2.

[113] Adam Rae, Vanesssa Murdock, Pavel Serdyukov, and Pascal Kelm. Working notes for the placing task at MediaEval 2011. In *Working Notes Proceedings of the MediaEval Workshop*, volume 807 of *CEUR Workshop Proceedings*, pages 1–2, 2011.

[114] Mark Sanderson and Yu Han. Search words and geography. In *Proceedings of the ACM Workshop on Geographic Information Retrieval*, GIR'07, pages 13–14, 2007.

[115] Diana Santos and Marcirio Silveira Chaves. The place of place in geographical IR. In *Proceedings of the ACM Workshop on Geographic Information Retrieval*, GIR'06, pages 5–8, August 2006.

[116] Frans Schalekamp and Anke Zuylen. Rank aggregation: Together were strong. In *Proceedings of the Workshop on Algorithm Engineering and Experiments*, ALENEX 2009, pages 38–51. SIAM, 2009.

[117] Steven Schockaert, Martine De Cock, and Etienne E. Kerre. Location approximation for local search services using natural language hints. *International Journal of Geographical Information Science*, 22(3):315, 2008.

[118] D. Sculley. Rank aggregation for similar items. In *SIAM International Conference on Data Mining*, SDM'2007, pages 587–592, 2007.

[119] Pavel Serdyukov, Vanessa Murdock, and Roelof van Zwol. Placing flickr photos on a map. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'09, pages 484–491, 2009.

[120] Josef Sivic and Andrew Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, ICCV'2003, pages 1470–1477 vol.2, 2003.

[121] Michele Trevisiol, Jonathan Delhumeau, Hervé Jégou, and Guillaume Gravier. How INRIA/IRISA identifies geographic location of a video. In Larson et al. [73], pages 1–2.

[122] Michele Trevisiol, Hervé Jégou, Jonathan Delhumeau, and Guillaume Gravier. Retrieving geo-location of videos with a divide & conquer hierarchical multimodal approach. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, ICMR'13, pages 1–8, 2013.

[123] Florian A. Twaroch, Philip D. Smart, and Christopher B. Jones. Mining the web to detect place names. In *Proceedings of the ACM Workshop on Geographic Information Retrieval*, GIR'08, pages 43–44, 2008.

[124] Koen E. A. van de Sande, Theo Gevers, and Cees G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[125] Jan C. van Gemert, Cor J. Veenman, Arnold W. M. Smeulders, and Jan-Mark Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283, 2010.

[126] Øyvind Vestavik. Geographic information retrieval: An overview. online, 2003. IDI, NTNU, Norway. http://www.idi.ntnu.no/õyvindve/article.pdf.

[127] Ville Viitaniemi and Jorma Laaksonen. Experiments on selection of codebooks for local image feature histograms. In *Visual Information Systems. Web-Based Visual Information Search and Management. Proceedings of International Conference on*

Visual Information Systems, volume 5188 of *Lecture Notes in Computer Science*, pages 126–137, 2008. VISUAL 2008.

[128] Zhao Xu, Xiaowei Xu, and Volker Tresp. A hybrid relevance-feedback approach to text retrieval. In *Advances in Information Retrieval. Proceedings of European Conference on Information Retrieval Research*, volume 2633 of *Lecture Notes in Computer Science*, pages 281–293, 2003. ECIR 2003.

[129] Seungwon Yang, Andrea L. Kavanaugh, Nádia P. Kozievitch, Lin Tzy Li, Venkat Srinivasan, Steven D. Sheetz, Travis Whalen, Donald Shoemaker, Ricardo da Silva Torres, and Edward A. Fox. CTRnet DL for disaster information services. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL'11, pages 437–438, 2011.

[130] H. Peyton Young. An axiomatization of borda's rule. *Journal of Economic Theory*, 9(1):43–52, 1974.

[131] Bo Yu and Guoray Cai. A query-aware document ranking method for geographic information retrieval. In *Proceedings of the ACM Workshop on Geographic Information Retrieval*, GIR'07, pages 49–54, 2007.

[132] Harry Zhang, Liangxiao Jiang, and Jiang Su. Augmenting naive bayes for ranking. In *Proceedings of the International Conference on Machine Learning*, ICML'05, pages 1020–1027, 2005.

[133] Xin Zhou, Adrien Depeursinge, and Henning Müller. Information fusion for combining visual and textual image retrieval in imageCLEF@ICPR. In *Recognizing Patterns in Signals, Speech, Images, and Videos. Proceedings of the International Conference on Pattern Recognition*, volume 6388 of *Lecture Notes in Computer Science*, pages 129–137, 2010. ICPR 2010.

[134] Alvaro Zubizarreta, Pablo de la Fuente, José M. Cantera, Mario Arias, Jorge Cabrero, Guido García, César Llamas, and Jesús Vegas. Extracting geographic context from the web: GeoReferencing in MyMoSe. In *Advances in Information Retrieval. Proceedings of European Conference on IR Research on Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, pages 554–561, 2009. ECIR 2009.